WHOLE-TRANSCRIPTOME ANALYSIS OF PROTEIN-CODING POTENTIAL IN THE MODEL PLANT *MEDICAGO TRUNCATULA*

by

Umut Çakır

B.S., Molecular Biology and Genetics, Boğaziçi University, 2020

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfilment of
the requirements for the degree of
Master of Science

Graduate Program in Molecular Biology and Genetics
Boğaziçi University
2022

This thesis is dedicated to my parents.

For their endless love, support, and encouragement

ACKNOWLEDGEMENTS

Foremost, I would like to thank my advisors Asst. Prof. Necla Birgül and Asst. Prof. Igor Kryvoruchko for his endless support. I am forever grateful for the encouragement and invaluable insight during my study. They are excellent mentors and friends who have been supportive throughout the processes, both in the research and in my professional life.

I would like to thank the committee members Assoc. Prof. N. C. Tolga Emre, Asst. Prof. Steven Footitt, and Prof. Bünyamin Akgül for their time and valuable feedback.

Many others have also been a source of inspiration, encouragement, and support, and without their help this would not have been possible. I'd like to give special thanks to Assoc. Prof. Marie Brunet and Prof. Xavier Roucou for their continuous advice, wisdom, and dedication from the very beginning of the project.

My deepest thanks to Assoc. Prof. Noujoud Gabed for her help in the field, providing information about data analysing in high-performance computing clusters and answering all my questions whenever needed.

Many thanks to Assoc. Prof. Harald Barsnes and GitHub community for their mentorship and invaluable guidance.

Mustafa Barbaros Düzgün and Yunus Emre Köroğlu, I have learned a lot from our discussions. I am thankful for your contributions.

Most notably, I would like to thank my family and beloved friend Yağmur Tarhana for their emotional support throughout all these years and for letting me follow my dreams.

This study was funded by the Scientific and Technological Research Council of Turkey (TUBITAK) 1002 Short Term R&D Funding Program (No. 120Z247) and Boğaziçi University Scientific Research Projects, BAP, Funding Program (No. 18841). I am grateful for the funding and hope that these programs can support many young

scientists in the future as well. I would like to thank TUBITAK ULAKBIM, High Performance and Grid Computing Center, TRUBA, for offering the opportunity to analyse a huge amount of data.

ABSTRACT

WHOLE-TRANSCRIPTOME ANALYSIS OF PROTEIN-CODING POTENTIAL IN THE MODEL PLANT *MEDICAGO TRUNCATULA*

How many different proteins can be produced from a single spliced transcript? Genome annotation projects usually do not consider the coding potential of altORFs. However, many altProts have been shown to carry out essential functions in various organisms. In addition to the existence of protein-coding potential in all the three reading frames, spliced eukaryotic transcripts may undergo programmed single or multiple ribosomal frameshifting events. Depending on whether a protein is produced by one or several such events, this novel protein is called either a chimeric protein or a mosaic protein, respectively. Proteins produced via single ribosomal frameshifting events have been known in viruses for a long time, and more recently, they have also been found in higher eukaryotes. In contrast, mosaic proteins so far are elusive, with only one example found in viruses. Detection of altORFs can help identify these unusual proteins because altORFs may act as building blocks for chimeric proteins and mosaic proteins. This way of extracting and combining genetic information from different reading frames may significantly increase proteome diversity, thus promoting organisms' flexibility and adaptability to various environmental conditions. This project aims to identify altProts based on the conservation evidence or detection by mass spectrometry (MS) analysis and to find proteins produced via single and multiple ribosomal frameshifting events to demonstrate the existence of mosaic translation. Our study in *Medicago truncatula*, a well-established model legume, detected 715 translated altProts and 146 chimeric proteins. Two transcripts support the existence of mosaic proteins and mosaic translation, which has never been detected in non-viral organisms before. In addition, we have found evidence for many thousands of conserved altProts. This work pioneers a new field of proteomics and is of immense value for plant biologists and specialists interested in translation. It also paves a way towards the major shift in current understanding of proteome complexity and diversity.

ÖZET

MEDICAGO TRUNCATULA MODEL BİTKİSİNİN PROTEİN KODLAMA POTANSİYELİNİN TAM TRANSKRİPTOM ANALİZİ

Tek bir kırpılmış transkriptten kaç farklı protein üretilebilir? Genom adlandırma projeleri genelde altORF'lerin kodlama potansiyelini dikkate almazken, birçok altProt'un çeşitli organizmalarda önemli işlevleri olduğu gösterilmiştir. Üç okuma çerçevesinin tümünde protein kodlama potansiyelinin varlığına ek olarak, kırpılmış ökaryotik transkriptler, programlanmış tekli veya çoklu ribozomal çerçeve kaymasına maruz kalabilir. Bir protein tek bir kırpılmış transkriptten tekli ya da çoklu ribozomal çerçeve kayması olaylarıyla üretilmesine göre, bu yeni proteinler sırasıyla kimerik ve mozaik protein olarak adlandırılır. Tekli ribozomal çerçeve kayması olaylarıyla üretilen proteinler virüslerde uzun süredir bilinmekte ve daha sonra yüksek ökaryotlarda da bulunmuştur. Buna karşılık, şimdiye kadar mozaik proteinlerin varlığının tarif edilmesi zordur ve virüslerde sadece bir örneği bulunmuştur. AltORF'lerin tespiti bu olağandışı proteinleri tanımlamak için kullanılabilir, çünkü altORF'ler, kimerik ve mozaik proteinler için yapı taşları olarak hareket edebilir. Farklı okuma çerçevelerinden gelen genetik bilginin bu şekilde ayıklanması ve birleştirilmesi, proteom çeşitliliğini önemli ölçüde artırabilir, organizmaların esnekliğini ve çeşitli çevresel koşullara karşı uyum sağlama kabiliyetini artırabilir. Bu proje, sekansların koruma kanıtlarına veya MS analizi ile tespite dayalı altProt'ları tanımlamayı ve mozaik translasyonun varlığını göstermek için tekli ve çoklu ribozomal çerçeve kaydırma olayları yoluyla üretilen proteinleri bulmayı amaçlamaktadır. Çalışmamızda iyi bilinen bir model legüm olan *Medicago truncatula*'da translasyona uğrayan 715 altProt, 146 kimerik protein tespit edilmiştir. İki transcript mozaik proteini ve daha önce viral olmayan organizmalarda hiç tespit edilmemiş olan mozaik translasyonun varlığını desteklemektedir. Ayrıca, binlerce korunmuş altProt tespit edilmiştir. Bu çalışma yeni bir proteomik alanına öncülük etmektedir ve bitki biyologları ve translasyon ile ilgilenen uzmanlar için büyük değer taşımaktadır. Ayrıca, proteom karmaşıklığı ve çeşitliliğine ilişkin mevcut anlayışta büyük bir değişimin yolunu açmaktadır.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS

$\bar{x}$	Mean
SD	Standard deviation
n	Number of samples
N	Number of modelled proteins
Q50	50% quartile
R	Row
v.	Version

LIST OF ACRONYMS/ABBREVIATIONS

aa	Amino acids
altORF	Alternative open reading frame
altProt	Alternative protein
APCI	Atmospheric pressure chemical ionization
CAS31	Cold acclimation-specific 31
CBS1	Cystathionine beta-synthase
CDS	Coding sequence
CID	Collision-induced dissociation
CNV	Copy number polymorphism
cRAP	Common repository of adventitious proteins
DART	Direct analysis in real-time
ESI	Electrospray ionization
FDR	False discovery rate
GWAS	Genome-wide association studies
hGSH	Homo-glutathione

hGSHSb	Homo-glutathione synthase b
HIV-1	Human immunodeficiency virus type 1
HTLV-1	Human T-lymphotropic virus type I
HTLV-2	Human T-cell leukaemia virus type II
INDELs	Inserts/Deletions
MATE67	Multidrug and toxic compound extrusion 67
MGF	Mascot generic format
MMTV	Mouse mammary tumor virus
M-PMV	Mason-pfizer monkey virus
MS	Mass spectrometry
MS/MS	Tandem mass spectrometry
N	Nitrogen
NCR169	Nodule-cysteine-rich protein 169
NCR247	Nodule-cysteine-rich protein 247
NH3	Ammonia
NO3-	Nitrate
NRamp1	Natural-resistance-associated macrophage protein 1
ORF	Open reading frame
PHO2-like	E2 ligase phosphate2-like
PTM	Post-translational modification

refORF	Reference open reading frame
refProt	Reference protein
REV1	Revoluta 1
RIBO-Seq	Ribosome profiling
RN	Row number
SCR	SCARECROW
SNF	Symbiotic nitrogen fixation
SNP	Single nucleotide polymorphism
WGS	Whole genome sequencing

INTRODUCTION

Nitrogen (N), a key component of nucleic acids and amino acids, is an essential nutrient source for plants, but its availability is limited in most soils. Although the atmosphere is composed mostly of dinitrogen gas N_2 (78.1%), plants cannot directly use the atmospheric N because of the very strong triple covalent bond between N atoms. Plants can absorb N in the reduced form, such as ammonia (NH_3) and nitrate (NO_3^-), which are produced from expensive and non-renewable energy sources. Thus, high-cost nitrogen fertilizers are used in modern agriculture to support the high yields necessary for a steadily growing world population (De Bruijn, 2019; Root-Bernstein & Root-Bernstein, 2016). However, using synthetic nitrogen fertilizers creates major environmental problems and will be impossible after the fossil energy sources are depleted (Sainju et al., 2020). Symbiotic nitrogen fixation (SNF) is the natural alternative to synthetic fertilizers. One of the very few plant lineages capable of SNF, legumes, are a vital nutritional source for animals and humans. They obtain N from the air and increase soil productivity by establishing a symbiosis with N-fixing soil bacteria called rhizobia. Specialized symbiotic organs called root nodules are formed on the roots of legume plants that undergo a symbiosis with rhizobia. Rhizobia can convert atmospheric N into NH_3 that can be metabolized by plants (Oldroyd et al., 2011). Understanding the molecular mechanism of the symbiotic relationship between legumes and rhizobia is necessary for world agriculture because it opens the possibility of engineering the symbiotic capacity in major non-legume crops such as wheat, rice, and maize (Santi et al., 2013). However, this mechanism cannot be resolved efficiently without a comprehensive characterization of genes involved in SNF. Current plant genome annotation projects focus exclusively on reference open reading frames (refORFs) as protein-coding units and do not consider the true coding capacity of annotated transcripts. This creates a considerable gap in our knowledge that is essential for the progress in legume molecular genetics and in plant genetics in general. Understanding the whole coding capacity and proteome complexity of plant genomes will accelerate crop improvement and the transfer of SNF to non-legume plants.

Alternative proteins (altProts), which are translation products of alternative open reading frames (altORFs), chimeric proteins, which are products of single ribosomal

frameshifting events, and mosaic proteins, which may result from multiple ribosomal frameshifting events, add a new dimension to proteome complexity. AltProts, chimeric proteins, and mosaic proteins can be referred to as a type of “the dark proteome” as their existence and functions are currently unknown. Identification of these novel proteins will help fully understand many fundamental biological processes in plants, including the mechanism of the symbiotic relationship of legume plants with rhizobia. Unravelling this domain of the dark proteome will shine a light on new ways to advance plant molecular biology, especially legume research necessary for sustainable agriculture and global food security.

In this study, *Medicago truncatula*, also called barrel medic (Watson et al., 2003), well known as a model legume in plant molecular biology research, was used to identify altProts, chimeric proteins, and mosaic proteins. *M. truncatula* is commonly used as a model organism because it has a rapid generation time, relatively small diploid (2n = 16) genome, prolific seed production, and autogamous nature (Ané et al., 2008; Kang et al., 2019). According to our other work in progress, more than 300 genes have been functionally characterized in *M. truncatula*, nearly two-thirds of which are involved in SNF. One of the goals of our work was to facilitate the accuracy of genetic studies in *M. truncatula* by detecting conserved and/or translated altORFs in all known transcripts of this organism.

ORFs are classically defined as spans of an RNA sequence between a translation initiation codon (start codon) and a translation termination codon (stop codon) within each reading frame of a transcript. Standard genome annotation projects annotate only the longest ORF in each transcript as a refORF if it begins with a start codon AUG (Kute et al., 2022). However, translation of many proteins is initiated with non-AUG start codons (Kearse & Wilusz, 2017). Because the whole range of codons that may initiate translation is unknown, our study defines ORFs as regions of an RNA sequence that do not contain a stop codon. According to this broad definition, these can be sequences at either end of the transcript delimited by the nearest stop codon or sequences between two stop codons anywhere in the transcript. Moreover, a functional translated ORF does not have to be the longest possible ORF in a transcript. For this reason, the most unbiased genome annotation pipeline called OpenProt does not use the artificial criterion of the longest length for

defining translated and potentially translated ORFs in genomes (Brunet et al., 2021). Instead, evidence for translation is obtained from direct approaches such as mass spectrometry (MS) proteomics and less direct approaches such as evolutionary conservation. In our work, we follow the principles and methodology of OpenProt with two exceptions. First, we extend our focus to ORFs that are shorter than 90 nucleotides in length (the shortest ORF considered in our approach is 60 nucleotides in length). It should be noted that even peptides as short as five amino acids in length can have vital biological functions (Yu et al., 2020, 2022). Second, for the reasons mentioned above, our detection pipeline is extended to ORFs that start with non-AUG codons. ORFs other than refORFs are called altORFs, and their translation products are called altProts. As we explained in our recent viewpoint article (Çakır et al., 2021), altORFs are likely to act as building blocks for chimeric proteins and mosaic proteins produced by ribosomes that shift from one reading frame to another in a programmed fashion as a response to internal and external stimuli. This expectation is based on the observation that many altProts share striking similarities to annotated proteins and is further supported by the analogy of this mode of translation to alternative splicing, which enabled eukaryotes to conquer so many ecological niches (Singh & Ahi, 2022). Therefore, we have proposed that mosaic translation must play a fundamental role in expanding the diversity of proteomes in all living forms.

Can we experimentally confirm the hypothesis of mosaic translation? This is a major technical challenge because current polypeptide detection methods cannot deliver a continuous amino acid sequence longer than 35 units (Meyer, 2014). Much longer polypeptide sequencing reads are necessary to detect chimeric proteins and mosaic proteins directly. Nevertheless, we have developed a strategy that enables the detection of such polypeptides even in the absence of long-read sequencing methods. Chimeric proteins are polypeptides which are comprised of two different portions, each derived from an overlapping ORF within one transcript. Examples of such polypeptides are known beyond the kingdom of viruses, in which they were discovered (Atkins et al., 2016; Dinman, 2006; Ketteler, 2012). They are produced when a single ribosomal frameshift changes the reading frame during translation without prematurely terminating the protein synthesis. Furthermore, when more than one ribosomal frameshifting event occurs during translation of a single transcript, a mosaic protein is produced. In contrast to chimeric proteins, apart

from the Gag-Pro-Pol polypeptide, mosaic proteins have never been detected, even though their existence was anticipated in literature (Ketteler, 2012). Identification of these novel proteins, which would be part of the dark proteome, will contribute to the uncovering of the underappreciated complexity of proteomes. No known attempt has been made so far to search for mosaic proteins, making this work very novel and unique. Regardless of the immediate success of this study, it paves the road towards the identification of such proteins in the future when more MS proteomic data become available for *M. truncatula*. Conserved and translated altORFs detected by us will serve as a starting point on this journey, which may ultimately lead to the discovery and quantification of mosaic translation first in *M. truncatula,* and then in other species since our protocol can be applied to any organism, including human.

1.1 Factors Contributing to the Proteome Complexity and Diversity

The genome is a very complex entity not limited to the multitude of all protein-coding genes. The array of transcripts produced from a given genome is much more diverse compared to the number of all transcribed regions. Furthermore, the array of polypeptides and proteins produced from those transcripts is by far more diverse compared to the sequences present in the transcriptome. Thus, the proteome is probably at least as complex as the genome and at the same time is much more complex and diverse compared to the transcriptome. This high-level complexity and diversity of the proteome emerge from variations in the processes that extract the information stored in DNA and interpret it for protein synthesis. The first process operates at the level of transcript maturation. Alternative splicing produces multiple transcripts from the same gene, which adds great flexibility to organisms in coping with environmental conditions and diversifies cell- and tissue-specific transcriptomes necessary for their specialized functions (Ren et al., 2021). Posttranscriptional changes to a mature transcript known as RNA editing can create a protein different from the normal amino acid sequence of its refORF. Both normal protein and its edited version can be present in the whole proteome at the same time regardless of whether they are translated in the same cell or different cells. This expands the number of proteins corresponding to a single gene even though its contribution is minor compared to alternative splicing (Farajollahi & Maas, 2010). Then, the proteome complexity and diversity are further enhanced at the level of translation and in the course of

posttranslational events. Cleavage of long precursor polyproteins into functional polypeptides, differential cleavage of normal proteins, alternative translation initiation sites, and post-translational modifications (PTMs) contribute to the elevated complexity and diversity of the proteome. Moreover, an additional layer of diversification is added at the protein complex level; protein subunits can assemble into multiple configurations, generating an array of complexes with various functionalities (Bludau & Aebersold, 2020). Additionally, splicing may occur not only on transcripts but also on proteins, known as peptide splicing. Peptide splicing by the proteasome is a post-process of translational splicing from a precursor protein. After translation, an internal protein segment is excised from the precursor protein, and the remaining fragments are ligated to form a novel protein by transpeptidation either in the sequential or reverse order (Vigneron et al., 2017). These proteins, produced by peptide splicing, may have a role in many cellular processes, such as cellular immunity. For example, they are presented at the cell surface by major histocompatibility complex class I molecules (Vigneron et al., 2017). Overall, peptide splicing by the proteasome contributes to proteome complexity and diversity.

Similar to the processes mentioned above, translation of altORFs that results in the synthesis of altProts, chimeric proteins, and mosaic proteins contributes to proteome complexity and diversity. The existence of some altProts and chimeric proteins is well established (Byun et al., 2005; Cardon et al., 2020; Fabre et al., 2022; P. Xu et al., 2018). However, such proteins are typically left beyond the scope of studies that use loss-of-function and gain-of-function approaches to deduce the biological roles of various genetic loci. Mosaic proteins are currently unknown. Nevertheless, their existence is anticipated and may prove to be essential for cellular processes. AltProts, chimeric proteins, and mosaic proteins should be identified to determine their biological roles and contribution to observed phenotypes. In the upcoming sections, altProts, chimeric proteins, and mosaic proteins are explained in greater detail.

1.2 *Medicago truncatula*, a Well-Established Model Legume Species

Medicago truncatula, also known as barrel medic (Watson et al., 2003) or barrel clover (Kong et al., 2021), is a model legume species for understanding plant-microbe interaction, seed development, and abiotic stress on plants (De Bruijn, 2019). *M.*

truncatula has a diploid (2n = 16) genome and is part of the *Medicago* genus, family Fabaceae, and subfamily Faboideae (Cook, 1999). Because it develops a symbiotic relationship with nitrogen-fixing rhizobia and arbuscular mycorrhizal fungi, it is commonly used for symbiosis research (Gavrin & Schornack, 2019).

Plants cannot directly use N from the air. Every N atom in the air is triple bonded to form molecular dinitrogen, N_2. The triple bond is strong. As a result, splitting the molecular N to obtain the raw atoms by a plant is energetically unfavourable. The process of breaking down the triple bonds between two atoms in a dinitrogen molecule is called nitrogen fixation (Mohapatra et al., 2014). Legume plants obtain N from soil bacteria called rhizobia, which fix N from the air. These microorganisms enter a symbiotic relationship with legumes and form nodules found on plant roots. They convert atmospheric molecular nitrogen N_2 to ammonia NH_3 or related nitrogenous compounds. In this way, the plants indirectly acquire nitrogen from the air through microorganisms. High-energy solar radiation and lightning can also split atmospheric molecular nitrogen, N_2. However, the amount of N fixed by these processes is insignificant compared to the amount fixed by microorganisms in the soil (Doane, 2017).

In legumes, N-fixing bacteria live in specialized symbiotic organs called root nodules or simply nodules. Plant roots release organic compounds as secondary metabolites called flavonoids that attract rhizobia to the root zone. Flavonoids then trigger the activation of *Nod* genes in the bacteria to synthesize Nod factors for initiating nodule formation. In response to these Nod factors, root hair curling begins. Curled root hairs capture individual rhizobia close to the root surface. Rhizobia encapsulated in a root hair cell proliferate and trigger transcriptional changes needed for nodule formation. Small tubes called infection threads are formed within the root hair cells, providing a way for rhizobial colonies to enter the epidermal cells of the root while the root hairs continue to curl. The movement of bacteria infects cells of the root cortex in the susceptible zone of the root, where rhizobia divide rapidly and are transformed into rod-shaped bacteroids. At this stage, bacteroids are surrounded by plant cell membranes to form structures called symbiosomes (Eckardt, 2006; Esseling et al., 2003; Limpens et al., 2005; Mergaert et al., 2020). Symbiosomes are N-exporting organelles of the infected nodule cells.

M. truncatula is used as a model plant organism in genetic and molecular analyses because it has many advantages: It has a rapid life cycle and is easy to maintain, modify, and breed in a laboratory setting. While it is a small legume species, it has a self-pollinating nature and the ability to produce a large number of seeds (Ané et al., 2008). These characteristics are very essential for practical reasons. Also, its relatively small genome (~375 Mbp) has been almost completely sequenced (Kang et al., 2019). The availability of a high-quality genome assembly and annotation makes it easier to study the molecular mechanisms and roles of various proteins, particularly those proteins involved in nodule formation and symbiosis.

M. truncatula research benefits from multiple genetic and genomic tools such as Gene Expression Atlas (MtGEA), which offers the opportunity to comprehensively compare the changes in gene expression during development in the main phases between the organs or conditions (Marzorati et al., 2021). In addition, large mutant populations based on physical, chemical, and insertional mutagenesis have been created in *M. truncatula*. The largest of them and the most actively used one consists of more than 22,000 insertional mutant lines (Kang et al., 2019). Besides, 384 sequenced inbred HapMap panels provide the basis for the detection of single nucleotide polymorphisms (SNPs), inserts/deletions (INDELs), and copy number polymorphisms (CNVs) between Medicago accessions at very high resolution and are useful for community-accessible genome-wide association studies (GWAS) (Cheng et al., 2022). According to the ScienceDirect® database, the number of studies on this model organism is increasing progressively, and about 5,500 articles that involve *M. truncatula* are published at the time of writing, 2022. Thus, we selected *M. truncatula* as a plant model to study altORFs and to find evidence for the existence of mosaic translation even though this species has less MS-based proteomic data compared to the human MS-proteomic resources (Deutsch et al., 2020). Our choice in favour of the plant model was dictated by the broader spectrum of functional genomics tools (for example, large mutant populations available), which are impossible in humans but are necessary for the downstream analysis of altProts, chimeric proteins, and mosaic proteins. Another rationale behind focusing on a legume rather than any other organism for studying altORFs, chimeric proteins, and for the proof of the mosaic translation hypothesis is our intention to advance genetic studies on SNF, for which *M. truncatula* is currently the best model.

1.3 Protein Validation by Mass Spectrometry

Proteomics is the study of all proteins in a biological system, such as cells, tissue, or organism, during specific biological events or conditions. In proteomics, mass spectrometry (MS) can be used to identify unknown proteins by molecular weight measurement, quantify known proteins, and determine the structure and chemistry of molecules, and it has become an increasingly important analytical technique for protein validation. MS measures ions' mass/charge ratio, shown as m/z, to identify and quantify molecules in the samples (Han et al., 2008).

An MS facility consists of at least the following three components: an ionization source, a mass analyzer, and an ion detector (Siuzdak, 2004). The sample is loaded into a mass spectrometer in liquid, gas, or dried form and then vaporized and ionized by an ion source such as atmospheric pressure chemical ionization (APCI), direct analysis in real-time (DART), or electrospray ionization (ESI) (McEwen & Larsen, 2009). Ions encounter electric or magnetic fields from mass spectrometers which deflect the paths of individual ions; thus, the ions are sorted and separated based on m/z. Mass analyzers are used to separate all analytes in a sample for global analysis. Alternatively, they can be used as a filter to deflect only specific ions to the detector (Smith & Thakur, 2010). Commonly used mass analyzers are orbitraps, time-of-flight, ion traps, and quadrupoles (Savaryn et al., 2016). Moreover, each type of mass analyzer has particular characteristics; thus, mass analyzers are selected based on separation resolution, operation speed, and other operational requirements. Furthermore, ions deflected by the mass analyzer hit the ion detectors that are electron multipliers or microchannel plates. The detector emits a cascade of electrons when each ion reaches the ion detector; thus, detection sensitivity is improved. The entire process is carried out under extreme vacuum conditions (10^{-6}-10^{-8} Torr) to remove neutral and contaminating non-sample ions and gas molecules, which may collide with sample ions, change the paths of ions, and generate non-sample signal (Glish & Burinsky, 2008; Vekey et al., 2011)

To improve the sensitivity of MS, two (or more) mass analyzers are used, and this method is called tandem mass spectrometry (MS/MS). The most straightforward MS/MS instrument consists of two mass analyzers in series connected by a chamber known as a

collision cell. Before mass analysis, firstly, the tertiary structure of proteins is disrupted for the easy access of proteases, and then, the enzymatic proteolysis by a protease digests the protein sample. Trypsin is the most used protease in proteomics and cleaves the C-terminus of lysine and arginine (Neagu et al., 2022). Once a sample is separated by chromatography, molecules in the sample are ionized, and the first spectrometer (stated as MS1) separates these ions according to the m/z. Ions from MS1 are denoted as precursor ions or parent ions. Ions with a specific m/z ratio from MS1 (precursor ions) are selected and split into smaller fragment ions by collision-induced dissociation (CID). Split ions are denoted as fragmented ions, product ions or daughter ions (Mittal, 2015). Moreover, CID is a fragmentation technique, and ions are accelerated by the electric potential to increase their kinetic energy and collide with neutral molecules, usually argon, nitrogen, or helium (Sleno & Volmer, 2004). Photodissociation, also known as photofragmentation, is another commonly used fragmentation technique in MS/MS. In photodissociation, chemical bonds are broken down by photons (Borsovszky et al., 2021). Fragments from MS1 are then introduced into a second mass spectrometer (MS2), which separates and detects the fragmented ions according to the m/z. The fragmentation step allows a typical mass spectrometer to identify and separate ions with similar m/z ratios (Büyükköroğlu et al., 2018). Note that the procedure explained here is called bottom-up proteomics; briefly, proteins are digested into peptides by a protease for MS/MS.

Another approach is called top-down proteomics. MS/MS analyses intact or whole proteins without prior digestion into peptides in top-down proteomics. The bottom-up approach is useful for identifying and quantifying proteins and PTMs, but it provides little information on the protein structure. However, top-down proteomics can provide information on the protein structure (Neagu et al., 2022). Compared to the top-down approach, the bottom-up strategy is a more robust high-throughput method for protein validation with better bioinformatics tools available at present. For this reason, the bottom-up approach is more appropriate for identifying novel proteins (Gregorich et al., 2014), such as altProts, chimeric proteins, and mosaic proteins.

Computational tools that analyse spectra from MS2 show m/z of fragmented ions to determine protein sequence using two sequencing approaches, *de novo* sequencing and protein database search. *De novo* sequencing uses MS/MS for direct analysis based on the

m/z of fragmented ions. Therefore, this approach can determine protein sequences that are not in the protein database and/or come from organisms with un-sequenced genomes (Medzihradszky & Chalkley, 2015; Yang et al., 2019). The most important advantage of *de novo* sequencing is that it does not require a protein search database (Muth & Renard, 2018). However, this approach requires higher-quality data and may produce more errors than the protein database search approach (L. He & Ma, 2010). The advantage of the protein database search is that it compares the experimental mass spectra of peptides with a database of theoretically computed peptide spectra and identifies the peptide in the database with the best match to the sequence of the experimental peptide (Kertesz-Farkas et al., 2012; Pevzner et al., 2001). The protein database search approach is considered more reliable than *de novo* sequencing because *de novo* approach identifies spectra from a "universal" database, and one spectrum may correspond to more than one peptide sequence due to highly similar masses of amino acids and PTMs (C. Xu & Ma, 2006). Additionally, very similar masses cause problems not only in *de novo* sequencing but also in the protein database search. For instance, isoleucine and leucine have the same mass, so they are generally considered to be indistinguishable (Xiao et al., 2016).

The protein database search is more reliable than *de novo* sequencing as using the search database limits the number of possible peptide sequences per spectrum, but the inclusion of too many proteins in the search database may cause database inflation. This may raise concerns about reliable and sensitive peptide validation. Spurious proteins in the search database can result in an underestimated false discovery rate (FDR) (H. Li et al., 2016). Furthermore, the number of peptides identified by MS decreases as the number of proteins used in the search database increases, resulting in false negatives, which causes to overlook translated proteins that may have vital functions in a cell or a pathway (Kumar et al., 2017). The MS search database should be carefully designed so that its size is not inflated (Kumar et al., 2017). The smallest database that contains all the translated sequences should be chosen for reliable and sensitive protein validation. Thus, cell- and condition-specific RNA-seq databases can be used to generate the protein search database. Translational products of transcripts not expressed in a particular cell or condition should not be included in the database. If possible, publicly available transcriptomic data for the cell of interest for a specific condition can be used to generate a three-frame translation of all expressed transcripts (Khitun & Slavoff, 2019). However, generating cell- and

condition-specific database is not possible in some situations, such as metaproteomics or translated altORF analysis (Chatterjee et al., 2016; Khitun & Slavoff, 2019).

Several bioinformatics tools and strategies have been developed to address protein database inflation, but each has advantages and disadvantages. There is no universally-accepted tool or technique applicable to all situations (Chatterjee et al., 2016; C. Chen et al., 2020; S. Kim & Pevzner, 2014; Leblanc & Brunet, 2020; Santos et al., 2022; Sticker et al., 2017). Furthermore, proteomic analysis using a large database may take an enormous time and is not memory-efficient (Beyter et al., 2018; W. Zhang & Zhao, 2013). In general, large protein search databases are subdivided into smaller data packages group. For a memory-efficient search, the large database, D, is split into arbitrary small files, d_i, and the spectral file M is also split into small spectra files, m_i. Each m_i spectral file is searched against each d_i file (Beyter et al., 2018). As an extension of this common practice, the whole spectral file M (without splitting) is searched against each d_i file. Then, identified proteins, also called validated proteins, from these searches are concatenated to generate a search database, d_c (subscript "c" stands for concatenated), for the second search. In this second round, the whole spectral file M (without splitting) is searched against the concatenated file d_c. The known proteins are always included in the search database and this strategy is called a two-step search approach throughout the thesis.

1.4 Reading Frames & Open Reading Frames

A reading frame groups three successive bases in a sequence of nucleotides in DNA or RNA molecules into non-overlapping triplets (Pienaar & Viljoen, 2008). There are three reading frames in one direction on a DNA molecule, and since DNA is double-stranded, any DNA sequence can be read in six different ways: three are in a forward strand, and the other three are in a reverse strand. Forward reading frames located in the forward strand are designated +1, +2, and +3. Similarly, reverse reading frames located in the reverse strand are designated -1, -2, and -3. Because a double-stranded DNA molecule has six reading frames, any nucleotide change at the DNA level can theoretically affect up to six polypeptides (Lin et al., 2014).

Open reading frames (ORFs) are defined as the significant length of DNA or RNA sequence that can start with any codon - start codon ATG does not have to be the first in the ORF and may be absent- and end with one of the three termination codons: TAA, TAG, or TGA (Andreev et al., 2022; X. Cao & Slavoff, 2020). In the ORF definition, the significant length is the minimum length of ORF and is nothing more than an arbitrary choice. For example, an ORF can be defined as a stop-free region at least 450 nt in length or even at least 60 nt in length (Claverie et al., 1997; Ladoukakis et al., 2011). Short polypeptides translated from small ORFs, usually called small peptides, e.g. 5 to 30 aa, may participate in many critical processes in the cell, such as gamete interaction and pollen tube growth during male-female crosstalk in plants (J. Zhang et al., 2021). Thus, the significant length should be carefully defined based on the research purpose. Moreover, although one reading frame in a transcript, in general, may have more than one ORF, there is also a possibility that one reading frame may have no ORF above the significant length threshold.

After splicing, mRNA transcripts become competent for translation and are called mature mRNAs. Mature transcripts are transported to the cytoplasm. The ribosome reads the nucleotide sequence in triplets in 5’ to 3’ direction to produce a polypeptide chain during translation. The RNA sequence has only three forward reading frames due to its single-stranded nature. In the spliced transcript, each forward reading frame may have zero or more ORFs. Even though all ORFs on a single transcript have the capacity to code polypeptides or proteins, genome annotation projects typically assign the longest ORF, called refORF or coding sequence (CDS), as protein coding ORF. (Raj et al., 2016; H. Xu et al., 2010). Accordingly, proteins that originate from ORFs other than the refORF are typically overlooked by genome annotation pipelines and consequently are generally neglected in functional studies (Brunet et al., 2021). This project aims to detect and validate conserved and/or translated altORFs, which are non-canonical ORFs different from the refORFs, at the whole-transcriptome level.

1.5 The Dark Proteome and Alternative Proteins

The proteome is an organism's complete set of proteins, including proteins that differ by PTMs (Wecker & Krzanowski, 2007). This term can also be used to describe the

assortment of proteins produced at a specific time in a particular tissue or cell type (Patade et al., 2018). Unfortunately, the whole proteome in the cell cannot be known, and the unknown portion is called the dark proteome. In other words, the dark proteome is the subset of proteins that exist and may have functions but escaped identification for technical reasons (size, shape, chemical properties etc., that are incompatible with the modern detection pipelines). Consequently, such proteins are excluded from functional studies even though many of them can have vital biological roles (Patade et al., 2018; Perdigão & Rosa, 2019).

Currently, the number or the portion of proteins that fall into the category of the dark proteome is unknown. There are some predictions, but these are not very reliable. For instance, about 10% of the proteins in our cells are unknown to scientists from one perspective, but from another perspective, about 90% of the proteome could be under the dark proteome definition. The difference in measuring the relative size of the dark proteome depends on what is meant by protein identification or validation (Laura Howes, 2022). While protein identification is only the discovery of its existence from one perspective, from another perspective, protein identification is to reveal its functions and structures in addition to its existence (Perdigão et al., 2015; Perdigão & Rosa, 2019). The human proteome project states that a protein counts as identified if scientists have evidence of a peptide that matches the protein sequence predicted from a particular gene (Baker et al., 2017). However, there is no one-to-one relationship between genes and proteins. Each gene can be expressed in divergent ways, and proteins can be modified after they are produced (Gerstein et al., 2007).

Over the last 50 years, scientists have uncovered a considerable amount of information about the community of proteins that constitute living systems. However, each step in this progress shows that scientists have more to learn. If a protein belongs to the dark proteome, it does not mean it is unimportant. Many proteins which are not known currently have a function in the cell. We should explore the dark proteome to understand the true complexity of cellular processes. With this new knowledge, it will be possible to find more efficient treatments for diseases, engineer crops with desired characteristics, and make many more discoveries useful for practical purposes. For this reason, new concepts

and tools should be developed to detect, characterize, and understand proteins in the dark proteome (Laura Howes, 2022).

AltORFs are all potentially protein-coding regions other than refORFs, whose existence is overlooked in conventional genome annotation pipelines (Vanderperre et al., 2013). These pipelines assume the protein-encoding genes have a minimal length threshold, around 150-300 codons (Deonier et al., 2005; Kute et al., 2022), and one transcript encodes only one protein. However, these initial assumptions are incorrect, and very small ORFs also encode functional proteins that have essential roles (Guerra-Almeida et al., 2021; Guerra-Almeida & Nunes-da-Fonseca, 2020; Orr et al., 2021). Furthermore, in addition to alternative splicing, RNA editing, peptide splicing by the proteasome, and PTMs of proteins, the array of polypeptides that can be produced from a single gene is diversified by the polycistronic nature of some eukaryotic transcripts, which encodes two or more proteins and was recognized only recently (Karginov et al., 2017; Mouilleron et al., 2016). In contrast to the polycistronic organization of prokaryotic transcripts, where structural ORFs are positioned in sequential order, in eukaryotes, these ORFs occupy the same genetic space (they overlap) because they are in different reading frames. It is also essential to point out that prokaryotes also benefit from this added layer of complexity. Their polycistronic transcripts also contain multiple overlapping ORFs with translation potential (Brunet et al., 2018). This principle, originally discovered in viruses, is not limited to one kingdom but is universal: from viruses to humans. Thus, one gene may encode more than one protein with an entirely different amino acid sequence, a refProt and altProts, which may have similar or independent functions (Renz et al., 2020; Von Arnim et al., 2014). Several altProts have been identified, and their potential functional roles in cellular mechanisms are characterized in many studies (Qin et al., 2018; M. Zhang et al., 2018; Zheng et al., 2019). Moreover, 195 altProts are discovered in K562 human cells. Of which, 76% are derived from non-annotated RNAs that are not found in the RefSeq database. Only 29% of those 195 altProts are initiated with the start codon AUG, and the remaining ones have non-canonical start codons (Ma et al., 2014). Astounding discoveries of altProts have shown that translated altORFs may be present upstream and downstream of the annotated ORFs, within annotated ORFs at a different reading frame, and on the RNAs previously considered as noncoding RNAs. Previously ignored altProts, a subset of

the dark proteome, have gained interest recently; their identification and characterisation have elucidated their role in the cell (Orr et al., 2021).

1.6 Chimeric and Mosaic Proteins: Shedding More Light on the Dark Proteome

Similar to altProts, chimeric proteins, and mosaic proteins can be part of the dark proteome. Their presence is usually ignored in the conventional genome annotation pipelines. Chimeric proteins are composed of translation products of two overlapping ORFs located on the same transcript. These ORFs are translated with a frameshift towards the end of the first frame so that the ribosomal frameshifting site corresponds to the fusion between products of two different ORFs. Chimeric proteins and ribosomal frameshifting are commonly observed in but are not limited to viruses. The well-known chimeric protein in viruses is the Gag-Pol protein which is produced by a -1 (minus one) ribosomal frameshift (Dinman et al., 1991; Ribas & Wickner, 1998). Human immunodeficiency virus type 1 (HIV-1) uses a -1 ribosomal frameshift as a part of its life cycle to synthesize the required ratio of the Gag and Gag-Pol polypeptides (Biswas et al., 2004). One striking chimeric protein discovery in prokaryotes is the production of two copper-related proteins from the same gene in *Escherichia coli*. The copper ion transporter CopA is produced without a ribosomal frameshift, and its chaperone is translated in *E.coli* from the same gene by a -1 ribosomal frameshift (Meydan et al., 2017). Furthermore, the mammalian antizyme-1 protein is an example of a chimeric protein in higher eukaryotes. Accumulation of polyamines triggers a +1 ribosomal frameshift event required for the synthesis of that antizyme (Atkins et al., 2016).

Mosaic proteins are conceptually similar to chimeric proteins but are produced by at least two ribosomal frameshift events per transcript instead of one. Mosaic proteins are hypothetical polypeptide sequences produced by the mosaic translation mechanism (Çakır et al., 2021). Earlier, Ketteler (2012) suggested the possibility of more than one frameshifting event per transcript (Ketteler, 2012). However, to the best of our knowledge, the presence of mosaic proteins has never been considered in novel protein identification projects because MS-based protein identification is commonly accomplished with the sequence database, and the inclusion of all theoretical mosaic proteins inflates the search

database drastically so that protein identification becomes impossible due to the loss of sensitivity (Tariq et al., 2021).

The only known mosaic protein example was observed in viruses, although it is not called a mosaic protein in the literature. This mosaic protein is the Gag-Pro-Pol polypeptide, the same as mentioned above to illustrate a chimeric protein. After the translation of Gag, a -1 ribosomal frameshift causes the ribosome to change the Gag reading frame to the Pro reading frame; then, another -1 frameshift changes from the Pro reading frame to the Pol reading frame. Thus, two -1 ribosomal frameshift events create the full-length Gag-Pro-Pol precursor protein. This double frameshift is observed only in some retroviruses, such as Mouse Mammary Tumor Virus (MMTV), Mason-Pfizer Monkey Virus (M-PMV), and Human T-Lymphotropic Virus type 1 (HTLV-1). However, Pro lies in the Pol reading frame in some lentiviruses and HIV-1 (Hatfield et al., 1992; Jacks, 1990). As far as we know, apart from the Gag-Pro-Pol polypeptide, no other example of more than one frameshift per transcript leading to the translation of a fusion polypeptide has been discovered so far.

About 60% of publicly available mass spectra are currently assigned to no known protein. Although some of the unassigned spectra could be artefacts attributed to the low quality of detection, a significant portion of such orphan mass spectra are considered to be of high quality (Pathan et al., 2017). Unassigned, exceptionally high-quality spectra may correspond to the unknown proteins or the dark proteome. Thus, it is necessary to include altProts, chimeric proteins, and mosaic proteins in the MS search database in order to reveal the identity of unassigned spectra. In addition, a new methodology or a pipeline should be developed to identify these novel proteins and incorporate their detection in all genome annotation projects. Several altProts have been identified, and their roles in the cellular processes have been researched in recent years. However, while a few studies identified and characterized chimeric proteins, e.g. mammalian antizyme-1 and the copper-related protein, currently, there are no reports on the identification and characterization of mosaic proteins. Our study was designed to fill in this gap in our knowledge.

1.7 The OpenProt Project: the Novel Genome Annotation Database for AltProts

More than one distinct protein can be translated from one spliced transcript by the use of different ORFs. Each spliced transcript can be translated unidirectionally from three reading frames, and multiple ORFs are available for translation. That is, each ORF on a spliced transcript has the capacity to encode proteins. The longest ORF in a transcript is generally considered a refORF, and all other ORFs are called altORFs, some of which may be translated to altProts (Brunet et al., 2021). Research on altProts is a rapidly growing area (Samandi et al., 2017). While translation of altProts in prokaryotes has been extensively characterized, altProts in eukaryotes have received due consideration only recently (Brunet et al., 2021).

A novel genome annotation database, the OpenProt database, annotates and expounds both refProts and altProts across 10 species, yeast and nine animals. This database provides supporting evidence for the existence of altProts, such as protein conservation deduced from the multi-species sequence alignments and translation measured by ribosome profiling (RIBO-Seq) and MS-proteomic techniques. For example, in human, the OpenProt pipeline (v. 1.3) identified about 650,000 ORFs longer than 90 base pairs, 450,000 (69%) of which are thought to be altORFs. Among those 450,000 altORFs, about 275,000 altORFs have evidence from at least one detection method: conservation, RIBO-Seq, or MS-proteomics. These experimentally supported altORFs have the following partition: about 240,000 have conservation evidence, 5,000 have RIBO-Seq derived translation evidence, and about 30,000 have protein evidence by MS. The number of altORFs identified in humans and other species demonstrates that many proteins are translated from altORFs along with or instead of the canonical refORFs (Brunet et al., 2021). Research on the whole spectrum of potentially functional altORFs and their translational products, altProts, is crucial for interpreting the protein-coding portion of the genome. Identification and characterization of altProts is the key to the complete understanding of fundamental cellular processes, interactions, diseases etc., and can help develop new treatments, find new candidate polypeptides for synthetic biologists, and uncover many mysteries in the cell (Nelde et al., 2022; Orr et al., 2021; Vanderperre et al., 2013). Unfortunately, while 10 species are available in the OpenProt database, no plant

species are included so far. Thus, a comprehensive analysis of altORFs in plant organisms is necessary, especially in legume plants (Brunet et al., 2021).

1.8 Mechanisms of Ribosomal Frameshifting

Different proteins can be produced from the same spliced transcript by a ribosomal frameshift, also called a programmed translational frameshift, during translation, thus contributing to the proteome complexity (Ketteler, 2012). AltProts may act as building blocks for chimeric proteins and mosaic proteins, which are produced via single and multiple ribosomal frameshifting events, respectively, from a single mature transcript. A ribosomal frameshift changes the reading frame during translation, e.g. from +1 to +2; thereby, multiple proteins can be produced from a single spliced transcript (Çakır et al., 2021). A frameshift can be triggered and controlled by sequence-dependent and sequence-independent mechanisms (Atkins et al., 2016; Brierley et al., 2010). Sequence-dependent mechanisms include the recognition of specific RNA sequences that tell the ribosome to slip and skip one or more nucleotides; the process is known as a forward ribosomal frameshift, or step back one or more nucleotides is known as a backward ribosomal frameshift. These specific RNA sequences are also known as slippery sequences (Çakır et al., 2021; Firth et al., 2012). Although some slippery sequences can be conserved across species and genera, many known slippery sequences are specific for particular species or virus types. For instance, in the Gag-Pol polyprotein synthesis described in HIV1, a slippery sequence responsible for a -1 ribosomal frameshift (slipping back one nucleotide) is the X_XXY_YYH motif, where the underlined symbol is the codon boundary for the ORF, XXX is any identical three nucleotides, YYY is AAA or UUU, and H is A, U or C (Biswas et al., 2004; Dinman et al., 2002). However, the slippery sequence for a +1 ribosomal frameshift (slipping forward one nucleotide) does not have the same motif. The UUU_CGX motif is a slippery sequence for a +1 ribosomal frameshift in plant amalgaviruses (Nibert et al., 2016). However, similar to the -1 ribosomal frameshift, there is no conserved slippery sequence for the +1 ribosomal frameshift across species. Some ribosomal frameshift sequences for +2 and -2 frameshifts are known, found in particular species, but usually viruses. However, defining a universal slippery sequence is impossible (J. Charon et al., 2016; Choi et al., 2003; Kartali et al., 2021; Pickett et al., 2011; Z. Xu et al., 2001).

Moreover, sequence-independent mechanisms also have a role in ribosomal frameshifting activity during translation. A ribosomal frameshift can be triggered by the presence of an RNA secondary structure (Bhatt et al., 2021). Pseudoknots are RNA secondary structures containing two or more stem-loop motifs in which half of one loop is intercalated between the two halves of another loop (Peselis & Serganov, 2014). The pseudoknot structure is thought to pause the ribosome during translation. That is, the pseudoknot structure of an mRNA molecule physically blocks the movement of the ribosome so that a ribosomal frameshift is triggered (Brierley et al., 2010). In MMTV, the pseudoknot structure promotes a -1 ribosomal frameshift, which is an example of a ribosomal frameshift event triggered by an RNA secondary structure (X. Chen et al., 1996). An mRNA pseudoknot's mechanical strength and frameshifting efficiency are also correlated. However, too strong pseudoknots may stop downstream translation (Hansen et al., 2007).

Furthermore, *cis* and *trans*-acting elements, which are small molecules, proteins, or nucleic acids, also trigger frameshifting activity during translation. For instance, the polyamine level in the cell is controlled by a +1 frameshift product involved in a negative feedback loop mechanism. This frameshift is triggered by polyamine levels to stimulate the production of an inhibitory enzyme (Atkins et al., 2016). Besides, frameshifting events cannot be exclusively governed by only one mechanism; sequence-dependent and sequence-independent mechanisms can simultaneously trigger or prevent a frameshift. For example, in Human T-Cell Leukaemia Virus type II (HTLV-2), the slippery sequence of the Gag-Pro protein junction can trigger a basal level of the ribosomal frameshift, which is enhanced by a pseudoknot structure (Kollmus et al., 1994). In summary, the detection of chimeric proteins and mosaic proteins cannot benefit from the conservation-based prediction of sequence-dependent features because of the lack of conservation. Therefore, in our study, we developed an approach that takes into account all possible positions of a frameshift between two ORFs and involves modelling of potential frameshifting products.

2. MATERIALS & METHODS

2.1 Identification of All Theoretical AltProts

Medicago truncatula genomic sequences and annotated features (v. 5.1.7) were obtained from the Medicago truncatula A17 r5.0 genome portal (Pecrix et al., 2018). Three-frame in silico translation of each transcript was conducted with an in-house Python script (see section 2.7 Data Availability). The script takes cDNA sequences for mRNA, ncRNA, rRNA, and tRNA as an input and gives the translation product of all possible ORFs between any two stop codons or at either end of a transcript. ORFs which are equal to or longer than 60 (≥60) nt were taken into consideration, while shorter ORFs (<60 nt) were eliminated from the analysis. For mRNA transcripts, refORFs were determined using canonical protein sequences of the organism and excluded from the analysis. ORFs were *in silico* translated into protein sequences using the standard genetic code. The input sequences for this script must be in fasta format, and the output file can be saved in either fasta or XML format. Fasta-formatted output files were used throughout the thesis. Each output sequence was automatically supplied with a unique identifier that contains the following information: genetic locus, the direction of ORF, reading frame, coordinates of ORF on its transcript (first nucleotide and last nucleotide), and length of ORF in nucleotides. Each element of the identifier is separated from other elements with an underline ("_") symbol. For instance, altProt MtrunA17_Chr4g1018210_3F_21-332_312 was generated from transcript MtrunA17_Chr4g1018210 using the third reading frame, the ORF start is at the 21st and the ORF stop is at the 332nd nucleotide, which gives the total length of 312 nucleotides. Note that symbols 1F, 2F, 3F correspond to reading frames +1, +2, and +3, respectively.

2.2 Comparing AltProt Sequences Against a Protein Reference Database

All altProts were aligned to sequences from the reference protein database UniProt (v. 2020_02) for similarity search by DIAMOND (v. 0.9.14) (Buchfink et al., 2021). The UniProt reference database was downloaded, and sequences were concatenated to generate a single fasta file. The generated single fasta file was used with "makedb" DIAMOND

command to build a search DIAMOND database. Sequence similarity searches were conducted with the following parameters: the maximum number of hits, “max-target-seqs”, was set to zero so that all hits per query are reported; the expect value, “evalue” was set to 10^{-3} (1e-3), and an option “more-sensitive” was used to enable higher sensitivity of the searches.

For each altProt, if a significant similarity was found, that is, an altProt returned hit(s), the following information was gathered in tabular data format: query sequence title, “qseqid”, subject sequence title, “stitle”, alignment length, “length”, percentage of identical matches, “pident”, percentage of positive scoring matches, “ppos”, query coverage per high scoring segment pair, “qcovhsp”, expect value, “evalue”, and the number of hits. The last value was calculated by an in-house script. If significant similarity was not found, the information columns were left empty. For the top hit analysis, the top hit was taken for each query, and the other hits were dropped from the study. However, the information on all hits was recorded for in-depth analyses of phylogenetic relationships between altProts and refProts to be conducted beyond this thesis.

2.3 AltProt Validation by Mass Spectrometry Searches

MS searches for all theoretical altProts were conducted using two publicly available datasets, PXD002692 (Marx et al., 2016) and PXD013606 (Shin et al., 2021). SearchGUI (v. 4.0.41) (Barsnes & Vaudel, 2018) and its partner tool Peptide Shaker (v. 2.0.33) (Vaudel et al., 2015) were employed for these searches. Datasets PXD002692 and PXD013606 were analysed independently. Dataset PXD002692 belongs to the first global proteomic blueprint of *M. truncatula* and its rhizobial endosymbiont. PXD013606 is a multi-species proteomic dataset generated to compare and contrast the protein levels across multiple plant species including *M. truncatula*. While the former dataset has nine organs/conditions, nodules (at three different time points: 10, 14, and 28 days after inoculation), buds, flowers, leaves, roots, seeds, and stems, the latter dataset has only one organ/condition called whole plant. Raw data from these two datasets were converted to Mascot Generic Format (MGF) file format via ThermoRawFileParser (v. 1.1.2) (Hulstaert et al., 2020).

X!Tandem, MS-GF+, OMSSA, and Comet search algorithms were used in all searches. The *M. truncatula* reference protein database called refProt and the contaminant database known as cRAP (common Repository of Adventitious Proteins) gathered from https://www.thegpm.org/crap/ were included in the search database in addition to altProts. Carbamidomethylation of C was set to fixed modification, and acetylation of protein N-term and oxidation of M were set to variable modifications. Precursor and fragment tolerance were set to 4.5 ppm and 20.0 ppm, respectively. A maximum of two missed cleavages were allowed. PSMs, peptides, and proteins were validated at 1% FDR using target/decoy hit distribution. The decoy dataset was created by *in silico* reversing target sequences. Parameters that are not stated here were kept in default settings (based on the version). To increase the confidence of altProt identification, the following principle was adopted: If a validated altProt was found either in the refProt or in the cRAP database, the altProt was always eliminated from the analysis pipeline even though such altProt could be indeed translated in the source genome. In other words, if a validated altProt sequence was found in either of the refProt and/or cRAP databases, the altProt was excluded from the analysis because of the ambiguity of its source. Besides, Peptide Shaker software has its own classification procedure, which labels the validated proteins as “Confident” or “Doubtful”. However, we decided to include all validated altProts in the protein report regardless of the classification by Peptide Shaker. This measure was necessary to decrease FNR, which is very important in our study. With relatively few validated altProts available, it would be imprudent to disregard altProts marked as “doubtful” by the software because many of them are likely to be genuine and thus valuable targets for the downstream functional analysis. Wet lab experiments, namely those outlined in section 4.9, should have the final say in the confidence status of these altProts. At the same time, results of the altProt classification by Peptide Shaker are available; please see section 2.7 Data Availability.

Because the inclusion of all mRNA-derived altProts causes the inflation of the search database (~1 million after the inclusion of refProts), a two-step MS search approach was used. In the first step, the altProt database was split into 10 arbitrary groups. Each group was used as a search database, and proteins validated in this first step were subsequently concatenated for the second search. In the second step, the concatenated proteins were searched one more time. Because there were 10 organs/conditions in datasets

PXD002692 and PXD013606 combined and 10 groups of the mRNA-derived altProt sequences, 100 MS searches were conducted in the first step of the two-step approach and 10 in the second step for these two datasets.

The two-step approach was not used for non-mRNA-derived altProts because their relatively small number does not inflate the search database. Thus, a regular MS search protocol was used for non-mRNA-derived altProts. MS searches for ncRNA, rRNA, and tRNA-derived altProts were conducted separately. Because there were 10 organs/conditions in total in datasets PXD002692 and PXD013606 combined, 10 MS searches were conducted for each group of non-mRNA-derived altProts. All validated altProts were recorded for further analysis. Note that the MS datasets were also searched independently; validated proteins from dataset PXD002692 were not combined with those from PXD013606 for the search database of the second step.

2.4 Modelling of Chimeric Proteins

Each MS-validated and conserved altProt has a corresponding altORF with coordinates on its transcript and locus information. Our chimeric protein modelling algorithm determines and *in silico* translates many possible ribosomal frameshifting events that may occur if an altORF overlaps with its refORF or other altORFs on the same transcript. It should be noted that we use the word combination “many possible … events” instead of “all possible … events” to emphasize that some situations were deliberately left beyond the scope of our analysis. The reason for this decision was the phenomenon of the search database inflation described in section 1.3 and section 2.3. In short, generating models for all theoretically possible chimeric proteins is technically feasible with our algorithm. However, it is counterproductive to include all the models into the analysis. This is because the larger the database of theoretically possible chimeric proteins the lower the chance to find confidently validated MS peptides supporting the fact of chimeric translation (Kumar et al., 2017; H. Li et al., 2016). As can be seen from the corresponding formulas later in this section, the absolute numbers of all theoretically possible chimeric proteins associated with a single pair of overlapping ORFs are too large for a confident bioinformatic analysis. Thus, we included only the simplest situations that would serve as the most convincing illustration of chimeric translation without inflating the MS search

database. This explains why the settings for the generation of chimeric models were not uniform for all cases but were tailored to three specific scenarios described below. Furthermore, those settings depended on the location of the frameshift relative to the involved ORFs (the 5'-end vs the 3'-end). According to the first scenario, an altORF can be completely embedded in its refORF or another long altORF without reaching one of the UTRs. To make it easy to understand for the reader, only altORFs overlapping with refORFs will be explained in this paragraph. In the second scenario, an altORF can partially overlap its refORF so that some portion of the altORF is located in one of the UTRs. In the third scenario, there is no overlap between an altORF and a refORF so that the altORF is located entirely in one of the UTRs. The last scenario addresses an unconventional situation since ribosomal frameshifting over three or more nucleotides is generally not considered. However, the last scenario elucidates whether the translation product of non-overlapping ORFs can be combined in a single continuous polypeptide by ribosomal frameshifting over longer distances than generally assumed. This means two ORFs joined in this fashion could belong either to different reading frames or to the same reading frame, in contrast to all other scenarios.

The chimeric protein modelling algorithm takes a particular region corresponding to an overlap between two ORFs and then generates many possible chimeric proteins. An example of chimeric protein modelling is shown in Figure 2.1. In this example, an altORF (light red) is within its refORF (yellow). In this most common situation, chimeric proteins can be modelled in two ways that correspond to a frameshift either from the refORF to the altORF or the other way around. In Figure 2.1, chimeric proteins modelled for a frameshift from the refORF to the altORF are shown. The refORF and the altORF are located in the third frame and the first frame, respectively. The ribosome can switch from the third frame to the first frame in two different frameshifts: +1 and -2 frameshifting events. There are 42 possible chimeric proteins that can theoretically correspond to the frameshift from the refORF to the altORF if the minimum size of a sequence contributed by either ORF in a chimeric protein is chosen to be 10 aa. The figure was created using Geneious software v7.1 (Biomatters). Note that frameshifting events that involve more than two nt at a time are not considered in this scenario. First, the chimeric protein modelling algorithm takes the region between the 390^{th} (10 aa upstream from the start of altORF) and the 510^{th} (30 aa downstream from the start of altORF) nucleotides. Then it proceeds in one-nucleotide steps

(we call iterations) from left to right, moving the frameshift position with each step until the last modelled protein contains 30 aa from the refORF and 10 aa from the altORF. For this example, there are 20 iterations and 21 modelled chimeric proteins (the first model corresponds to iteration zero). The same principle, with some variations, was applied for other scenarios (Figure 2.2).

While the iteration procedure starts at the beginning of altORF for the left side of the first scenario, it starts at the beginning of refORF and altORF for the second scenario's 5’ and 3’ cases, respectively. Furthermore, in contrast to other cases, the modelling starts 30 aa upstream from the end of altORF for the right side of the first scenario. The first model produced in this case contains 10 aa from the altORF (not the last ones) and 30 aa from the refORF, while the last model contains 30 aa (the last ones) from the altORF and 10 aa from the refORF. The minimal number of 10 aa included from the first ORF in an overlapping pair is a setting common for scenarios one and two. Concerning the contribution from the second ORF in the overlapping pair, the minimal number of 10 aa included from this ORF is a setting valid only for embedded altORFs (scenario one). Namely, for both situations of scenario two, the last modelled protein contains 39 aa from the first ORF and only one aa from the second ORF in an overlapping pair. This setting was necessary to capture situations similar to the one described for the copper-related protein in *Escherichia coli* where a frameshift product contained a single amino acid from the alternative frame at the 3’-end of the chimeric protein (Meydan et al., 2017). For the third scenario (no overlap), the minimum size of a sequence contributed by each ORF in a chimeric protein was set to 20 aa, which means only frameshifting events that join two non-overlapping ORFs were considered. The overall length of each chimeric protein fragment limited to only 40 aa is a parameter common for all the three scenarios. These settings were chosen based on the range of length typical for MS-derived peptides (7-35 aa, (Swaney et al., 2010)). The idea was to make sure an MS peptide spans the frameshift position rather than matches either part of the chimeric protein. This, however, does not mean we anticipate the chimeric proteins to be of that short length only. As we hypothesize, they can be long molecules, but we need to focus on the short fragment corresponding to the frameshift site in order to validate their chimeric nature.

For the left side of the first scenario, the number (N) of modelled chimeric proteins corresponding to the shift from the refORF to the altORF can be calculated by the following formula: N = (overlap length – minimum size + 1) * 2. Thus, 42 chimeric proteins are modelled for the shift from the refORF to the altORF if the altORF is 90 nt long (30 aa) and if the minimum size of the altORF or the refORF part considered in the modelled chimeric protein is chosen as 10 aa. In this case, the overlap length corresponds to the whole length of the embedded altORF in amino acids.

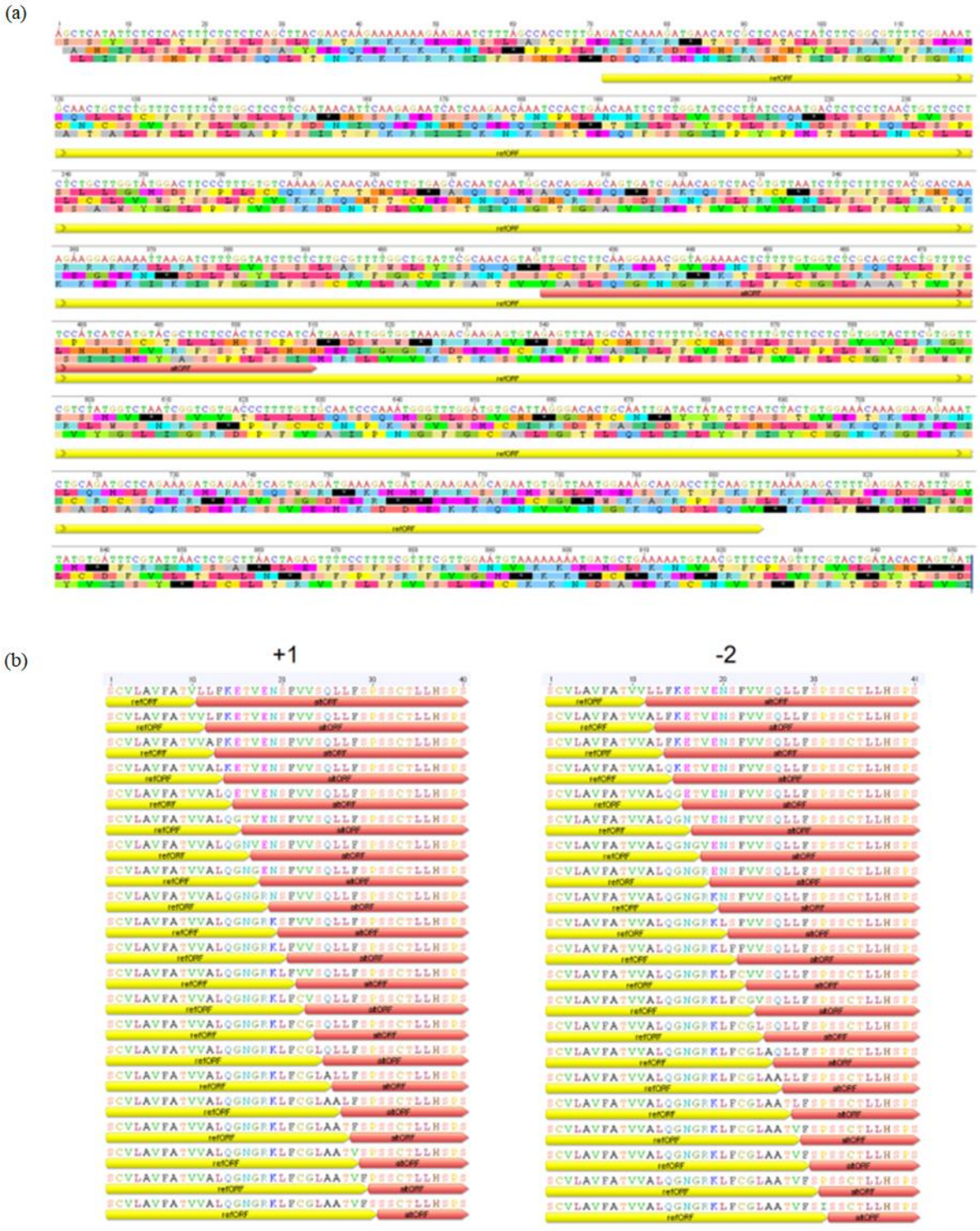

Figure 2.1. Modelling of hypothetical chimeric protein example. The upper panel (b) shows a hypothetical transcript and its three-frame translation. Two overlapping ORFs are present in the transcript: one is a refORF shown with yellow colour in the third frame and another is an altORF shown with light red in the first frame. The lower panel (b) shows many possible chimeric proteins produced by +1 and -2 ribosomal frameshifting events.

Below we describe further details of each scenario, corresponding settings, and the calculation formulas. The basis for the chimeric protein modelling algorithm is visualized in Figure 2.2. Black lines represent altORFs, and numbers on the black lines correspond to types of altORFs based on a position relative to the refORF: (1) represents altORFs that are within the refORF; (2) and (3) correspond to altORFs that overlap either the 5'-UTR or the 3'-UTR; (4) and (5) show altORFs that are located entirely either in the 5'-UTR or in the 3'-UTR without overlapping the refORF even by a single nucleotide, and (6) shows altORFs that span the whole refORF. According to Figure 2.2, position k corresponds to the beginning of refORF, positions l and m correspond to the beginning and end of altORF denoted with the number 1, and position n corresponds to the beginning of altORF denoted with the number 3. Note that the algorithm focused on a meaningful subset of all possible situations. For example, starting the iteration process at the end of altORF number 2, the end of refORF, or the middle of altORF number 1 was not considered, which helped alleviate the search database inflation problem.

For the type 1 altORFs, chimeric proteins are modelled separately at the 5'-portion and the 3'-portion of the altORFs. At the 5'-end of the altORF, the algorithm starts modelling 10 aa upstream from position l and proceeds until position l + 30 aa in 21 iterations (separate models). All corresponding chimeric proteins in this region are generated: the first chimeric protein is composed of 10 aa from the refORF and 30 aa from the altORF, and the last one is composed of 30 aa from the refORF and 10 aa from the altORF. In contrast, at the 3'-side of the altORF, which is used for the modelling of the switch to the refORF, 30 aa upstream from position m is taken and extended to position m + 10 aa in 21 iterations. All chimeric proteins in this region are generated: the first chimeric protein is composed of 10 aa from the altProt and 30 aa from the refProt, and the last chimeric protein is composed of 30 aa from the altProt and 10 aa from the refProt. The first and last 10 aa of all generated chimeric proteins are the same. Chimeric proteins are generated from the start of altProt and the end of altProt independently so that the total number of chimeric proteins is 21*2*2=84. In general, the formula that shows the number of chimeric proteins in (1) is

$$((overlap\ length\ -\ 10\ +\ 1) * 2 * 2) \tag{2.1}$$

where the first “2” comes from two frameshifting alternatives: forward frameshifting, e.g., a +1 frameshift, and back frameshifting, e.g., a -1 frameshift. Additionally, in Equation (2.1), the second “2” comes from consideration of chimeric proteins separately at the 5’ and 3’-sides of the altORF.

When an altORF overlaps its refORF at the 5’-side of the refORF (2), chimeric proteins are modelled in the following way: 10 aa upstream from position k is taken and extended to position k + 30 aa. All chimeric proteins in this region are generated: the first chimeric protein is composed of 10 aa from the altProt and 30 aa from the refProt. Similarly, the last chimeric protein is composed of 39 aa from the altProt and only 1 aa from the refProt. All generated chimeric proteins have the same first 10 aa; thus, the number of chimeric proteins is 30*2=60, where “30” indicates the number of overlapped regions used in the modelling algorithm, and “2” refers to the separate consideration of frameshifting events in the forward and in the backward direction.

When an altORF overlaps its refORF at the 3’-side of the refORF (3), chimeric proteins are modelled in the following way: 10 aa upstream from position n is taken and extended to position of n + 30 aa. All chimeric proteins in this region are generated: the first chimeric protein is composed of 10 aa from refProt and 30 aa from altProt; similarly, the last chimeric protein is composed of 39 aa from refProt and one aa from altProt. Similar to (2), all generated chimeric proteins have the same first 10 aa; thus, the number of chimeric proteins is 30*2=60, where “30” indicates the number of iterations (the first model corresponds to iteration zero), and “2” refers to forward and backward frameshifting events.

When an altORF does not overlap its refORF, but the gap between the altORF and the refORF is equal to or less than 10 nt (4) and (5), chimeric proteins are modelled in the following way: if the altORF is located at the 5’ UTR of the refORF (4), the last 20 aa of the altProt and the first 20 aa of the refProt are joined in sequential order. Furthermore, if an altORF is located at the 3’ UTR of the refORF (5), the last 20 aa of the refORF and the first 20 aa of the altProt are joined in sequential order. The number of modelled chimeric proteins is limited to one per situation.

In addition to the scenarios mentioned above, an altORF may span the whole refORF; that is, the beginning and end of the altORF are located at the 5’ UTR and the 3’ UTR of the refORF, respectively. In this situation, the beginning of altORF is assumed as in (2); that is, 10 aa upstream from position k (the beginning of overlap region) is taken and extended to position of k + 30. While the first chimeric protein is composed of 10 aa from the altProt and 30 aa from the refProt, the last chimeric protein is composed of 39 aa from altProt and one aa from refProt. Furthermore, the end of altORF is assumed almost as in (3) with the difference that 30 aa is taken upstream from the end of refORF (not from position n) and extended to the 40 aa. The first chimeric protein is generated as 10 aa from the refProt and 30 aa from the altProt; similarly, the last chimeric protein is composed of 39 aa from the refProt and one aa from the altProt.

In rare cases, chimeric proteins cannot be modelled as explained above. For instance, if an altORF is less than 60 nt and embedded in its refORF or there is no 10 aa sequence upstream from position k, the length of modelled chimeric proteins becomes less than 40 aa. In those types of rare cases, chimeric proteins may be modelled with overall length below 40 aa. Besides, although here we explained in detail the overlapping of an altORF with its refORF, an altORF overlaps with another altORF. In these cases, one altORF was assumed as a refORF, and many possible chimeric proteins were generated with the same procedure. Furthermore, if more than two ORFs overlap, all combinations of two ORFs were modelled, and possible chimeric proteins were generated with the same procedure. For instance, if two altORFs (A, B) and one refORFs (C) overlap on the same transcript, the following chimeric proteins were generated: A & C, B & C, A & B (ampersand symbol, &, indicates the composition of the modelled chimeric proteins). A & C, B & C are chimeric proteins composed of amino acid sequences contributed by the altORF and the refORF, and A & B are chimeric proteins composed of translational products of two altORFs. Furthermore, chimeric proteins were modelled from two lists of altProts. The first list contained altProts validated by MS searches, and the second list contained altProts with a minimum 70% top hit identity with an annotated protein. The first and second lists are called MS-validated and conserved altProts, respectively.

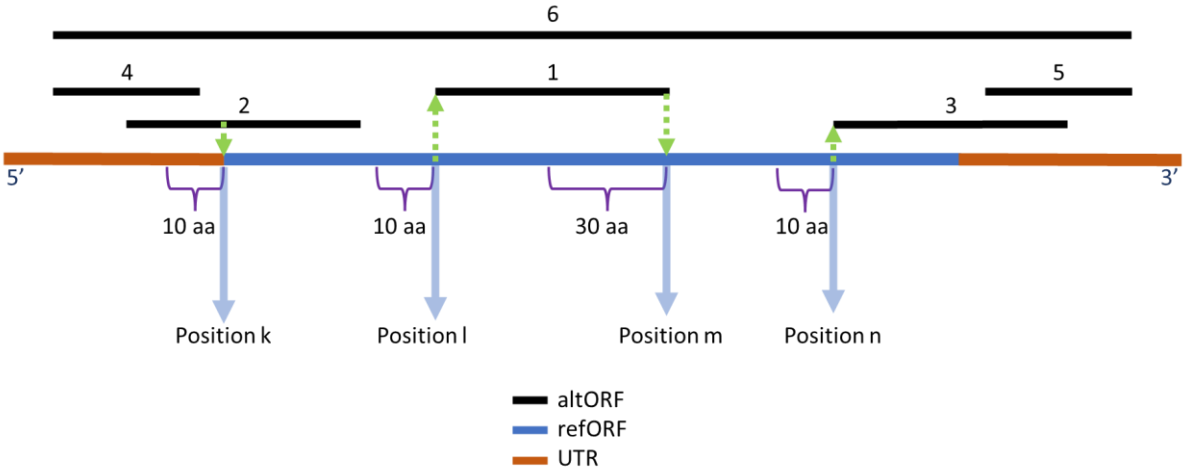

Figure 2.2. Graphical summary of the algorithm for chimeric protein modelling. AltORFs and a refORF are depicted with black and blue horizontal lines, respectively, and the UTRs of the refORF's transcript are depicted in orange. AltORFs can be located in six different places relative to their refORF.

2.5 Chimeric Protein Validation by Mass-Spectrometry Searches

Chimeric proteins were modelled from ORFs of altProts that are validated by MS searches (MS-validated altProts) and separately from altProts with the top hit % identity at least 70% (conserved altProts). While chimeric proteins modelled with MS-validated altProts do not inflate the search database, chimeric proteins modelled with conserved altProts inflate the search database drastically. While the number of chimeric proteins modelled with MS-validated altProts is ~32,000, it is ~530,000 for conserved altProts. Thus, similar to the MS-validation of mRNA-derived altProts, chimeric proteins modelled with conserved altProts were searched by the two-step MS approach. In contrast, because chimeric proteins modelled with MS-validated altProts do not inflate the database too much, the two-step MS approach was not necessary for that group. They were analyzed by regular MS searches.

Modelled chimeric proteins were searched in the same datasets, PXD002692 and PXD013606, by using SearchGUI (v. 4.0.41) (Barsnes & Vaudel, 2018) and PeptideShaker (v. 2.0.33) software (Vaudel et al., 2015). These two datasets together have 10

organs/conditions: nodules (at three different time points: 10, 14, and 28 days after inoculation), buds, flowers, leaves, roots, seeds, stems, and the whole plant. The order of chimeric proteins modelled with the conserved altProts database was shuffled, and the whole list was split into 10 equal groups. Each group was searched separately in 10 organs/conditions in the first step. Then, validated chimeric proteins were concatenated. Afterwards, concatenated validated proteins were searched one more time in the second step and validated proteins were recorded for further analysis. Because there were 10 organs/conditions in these two datasets, 100 and 10 MS searches were conducted in the first and second steps of the two-step approach, respectively. On the other hand, chimeric proteins modelled with MS-validated altProts were analyzed by regular MS searches and validated proteins were recorded for further analyses, especially for the analysis aimed to find evidence for mosaic translation.

The same search parameters that were used for the altProt validation were applied for the chimeric protein validation. Shortly, X!Tandem, MS-GF+, OMSSA, and Comet search algorithms were used in all searches. Carbamidomethylation of C was set to fixed modification, and acetylation of protein N-term and oxidation of M were set to variable modifications. Precursor and fragment tolerance were set to 4.5 ppm and 20.0 ppm, respectively. A maximum of two missed cleavages were allowed, and PSMs, peptides, and proteins were validated at 1% FDR using target/decoy hit distribution (decoy: reversed target protein sequences).

Furthermore, if a chimeric protein also exactly matches an altProt, a refProt, and/or an entry in the cRAP database, this chimeric protein is considered as not validated and is eliminated from the analysis pipeline. Thus, unlike in the altProt validation, altProts used for modelling chimeric proteins were also included in the search database in addition to the refProt and cRAP databases. The reason why altProts were included in the search database is that some chimeric proteins were similar to their altProts, such as different by only one or two aa different. If these different amino acids are indistinguishable by MS, true altProt may be categorized as a chimeric protein. Furthermore, in the two-step approach, although PXD002692 and PXD013606 datasets were searched independently for altProts, validated chimeric proteins from both databases in the first step were concatenated, and the same search database was used for these two datasets in the second step. In other words, if

chimeric proteins modelled with conserved altProts were validated using dataset PXD002692 in the first step, in the second step, they were searched using dataset PXD013606 or vice versa. The reason is that the number of validated chimeric proteins from the first searches does not inflate the database drastically. If many datasets would be used for chimeric protein validation, validated chimeric proteins from different datasets should not be concatenated for the second step.

2.6 Validation of Mosaic Proteins

AltProts can be building blocks for chimeric proteins and mosaic proteins. Translated altORFs, altProts, were, firstly, determined to identify mosaic proteins and to attest to the mosaic translation hypothesis. AltProts were identified by MS searches and conservation evidence. Then, chimeric proteins were modelled with altProts and validated by MS searches. After the validation of chimeric proteins, transcripts associated with chimeric proteins that received evidence for translation by MS searches were categorized into two groups. The first group encompassed transcripts with only one associated chimeric protein per transcript. The second group included transcripts that gave rise to at least two chimeric proteins. As the first group corresponds to chimeric proteins because only a single ribosomal frameshift event was proved, the second group corresponds to the only candidates for mosaic proteins because multiple ribosomal frameshifting events were demonstrated. Mosaic protein figures were generated using Geneious (v. 7.1) created by Biomatters and available from http://www.geneious.com.

2.7 Data Availability

All generated data from this study are available publicly in the Zenodo repository with the identifier: doi.org/10.5281/zenodo.7030093. All altProts and modelled chimeric proteins, DIAMOND outputs, MS search databases, and certificate of analysis and protein report files from all MS searches are available. Additionally, characterized genes for which conserved altProts and MS-supported altProts were found in this study are available in the Zenodo repository.

3. RESULTS

3.1 Identification of All Theoretical ORFs, AltProts, and RefProts

The minimum length threshold for altORFs in our analysis was set to 60 nt. AltORFs for mRNA, ncRNA, rRNA, and tRNA transcripts were determined separately. The number of transcripts, the median length of transcripts, number of altORFs, altORFs per transcripts, total number of nucleotides, and median length of altORFs or refORFs for each transcript group are shown in Table 3.1. RefORFs were determined and subtracted from all ORFs so that only altORFs were subjected to further analysis. AltProt and refProt sequences were generated by *in silico* translating altORF and refORF nucleotide sequences to protein sequences, respectively, using standard genetic code. The first column (mRNA^{b}) shows statistics before the elimination of refORFs; the second column (mRNA^{a}) shows statistics after the elimination of refORFs. In this case, a and b characters written as superscripts stand for “after elimination” and “before elimination”, respectively.

Groups of mRNA, ncRNA, rRNA, and tRNA transcripts contain 44,624, 5,657, 62, and 974 transcripts, respectively. Median transcript lengths were highest for mRNA (1,280 nt), followed by ncRNA (413 nt) and rRNA (120 nt), while the lowest value belongs to tRNA (75 nt). In total, ~875,000 theoretical altORFs are identified, with the majority belonging to mRNAs (~800,000) and the next largest group to ncRNA (~70,000). While mRNA has 18 altORFs per transcript, ncRNA and rRNA have 13 altORFs per transcript, and tRNA has only one altORF per transcript. The median length of refORFs (810) is nine-fold longer than the lengths of other ORFs, which illustrates why genome annotation projects typically annotate the longest ORF as a protein-coding ORF. While mRNA (90), ncRNA (93), and rRNA (96) have similar median lengths, tRNA has a relatively smaller median length, probably due to the smaller transcript length. Interestingly, despite the nearly 3.5-fold smaller median length of rRNA transcripts compared to ncRNA transcripts, rRNA has the same number of altORFs per transcript.

Table 3.1. Statistics of all theoretical altORFs grouped by RNA types

	mRNA[b] [1]	**mRNA[a] [2]**	**ncRNA**	**rRNA**	**tRNA**
Number of transcripts (refORFs)	44,624	44,624	5,657	62	974
Median length of transcripts, nt	1,280	1,280	413	120	75
Number of altORFs	846,711	802,087	71,127	831	1,311
AltORFs per transcript	19	18	13	13	1
Total nt	132,065,895	85,361,370	7,740,336	97,470	93,618
Median length of altORFs, nt	93	90	93	96	72
Median length of refORFs, nt		810			

[1] Before the elimination of refORFs
[2] After the elimination of refORFs

3.2 Conservation Evidence: AltProts with Similarity to at Least One Annotated Protein

DIAMOND software (v. 0.9.14) was used to compare all altProts with the UniProt reference protein database (v. 2022_01). In the search, the e-value was set to 0.001, and the "more-sensitive" option was used. The top hit derived from the sorting by score, which was a default option, was recorded for each query as conservation evidence. For robust conservation evidence for altProts, chimeric proteins, and mosaic proteins, only queries that have the top hit % identity equal to or higher than 70% with an annotated protein were selected.

AltProts that have at least one hit without the 70% identity threshold are summarised in Table 3.2. Then, 18,600 altProts have at least one hit among all theoretical altProts (875,356), which corresponds to ca. 2%. About 13,400 mRNA-derived and 4,400 ncRNA-derived altProts have at least one hit, which corresponds to about 2% and 6% of all theoretical mRNA-derived and ncRNA-derived altProts, respectively. On the other hand, about 390 rRNA-derived and 430 tRNA-derived altProts have at least one hit, and those numbers correspond to ~47% and 32% of all theoretical rRNA-derived and tRNA-derived

altProts, respectively. Note the profound disequilibrium in these proportions, which is highly unexpected under the assumption of no protein-coding capacity in non-mRNA transcripts.

On average, three in 10 mRNA transcripts have altProts with at least one hit to the reference database. This number is higher for the other transcript groups. For ncRNA, eight in 10 transcripts, and for tRNA, four in 10 transcripts have altProts with at least one hit. Interestingly, each rRNA transcript has six altProts with at least one hit. In other words, although every mRNA, ncRNA, and tRNA transcript has less than one altProt with significant hits, the rRNA transcripts clearly stand out according to this parameter. Together with the highest proportion of the scoring transcripts (47%) out of the total number of rRNA transcripts (see above), this could be evidence for the RNA world hypothesis that RNA with genetic information and catalytic activity that could copy itself without help from other molecules was essential in the origin of life (Saito, 2022).

The median lengths of mRNA, ncRNA, and rRNA-derived altProts with significant hits are 59, 50, and 45 aa, respectively, although the median length of all theoretical altORFs for the respective transcript groups (mRNA: 90, ncRNA: 93, rRNA: 96 nt, see Table 3.1) are in a closer range. In contrast, because the median length of all theoretical tRNA-derived altORFs is the smallest (tRNA: 72 nt, see Table 3.1), the median length of tRNA-derived altProts with significant hits is 25 aa, which is the least among all RNA groups. Similarly, the median length of alignment is somewhat close for mRNA (50 aa), rRNA (48 aa), and rRNA queries (45 aa), but it is much shorter for tRNA (25 aa), which is expected since tRNA-derived altORFs generate the shortest altProts.

The median % identity values of mRNA, ncRNA, rRNA, and tRNA-derived altProts with significant hits are 77, 84, 92, and 97, respectively. Similar to the % identity, the median % coverage of mRNA, ncRNA, rRNA, and tRNA-derived altProts with significant hits are 79, 80, 91, and 98, respectively. It should be noted that these parameters have the inverse relationship with the median length of transcripts in each group but not with the median length of altORFs, which may be another piece of evidence supporting the RNA world hypothesis and the special role of tRNA in the evolution of proto-genomes (Root-Bernstein & Root-Bernstein, 2016).

Note that row numbers (R denotes row) in the table are shown in the first column and are meant to show the reader how R4-R6 rows are calculated. Additionally, the numbers in the text are approximate, and absolute numbers are shown in the table.

Table 3.2. Summary of altProts with at least one hit in the global BLASTP analysis

#		**mRNA**	**ncRNA**	**rRNA**	**tRNA**
R1	Number of transcripts	44,624	5,657	62	974
R2	Number of altProts, regardless of BLASTP results	802,087	71,127	831	1,311
R3	Total number of altProts that have at least one hit	13427	4398	392	426
R4	Percentage of altProts that have at least one hit per all theoretical altProts in the group (R3/R2*100)	1.67	6.18	47.17	32.49
R5	AltProts per transcript (R3/R1)	0.3	0.78	6.32	0.44
R6	Transcripts per altProts (R1/R3)	3.32	1.29	0.16	2.29
R7	Median length of altProts, aa	59	50	45	25
R8	Median length of alignment, aa	50.2	47.8	44.8	24.5
R9	Median % identity	77	83.8	91.6	97.2
R10	Median % coverage (query coverage)	78.6	80.3	91.2	97.9

Another parameter potentially useful for studying the origin and functions of altORFs is the number of hits associated with each query, which we call frequency in Figure 3. This figure shows the frequency distribution of the number of hits per altProts and indicates how “popular” a protein sequence is in the tree of life and may serve as a proxy for the evolutionary age of a sequence (Malhis et al., 2019; Vanderperre et al., 2013). The range of the number of hits per query is very broad, as shown in Figure 3.1. The median of the number of hits is two, but the mean is 240, indicating some altProts have numerous hits; thus, the standard deviation is very high. So, while most altProts with hits have less than five hits, some altProts have more than 100 hits. Among ~18,600 altProts with hits, ~13,000 altProts (70%) have five or fewer hits. On the other hand, ~2,100 altProts (11%) have at least 100 hits.

Next, hits were then sorted by the score, which is the default sorting method of DIAMOND software, and the top hit for each altProt was selected for subsequent analysis.

Figure 3.2 and Figure 3.3 show the top hits % identity of altProts by overall and categorized styles, respectively.

The mean % identity of all altProts with hits is ~80%, which means most alignments in our dataset are very high % identity alignments. As most altProts with hits belong to mRNA and ncRNA groups, the mean values for mRNA- and ncRNA-derived altProts (~77% and ~84%, respectively) are closer to the mean of all altProts in terms of % identity. In contrast, the mean % identity values for rRNA and tRNA-derived altProts are higher at ~92% and ~97%, respectively. Although the mRNA and ncRNA or rRNA and tRNA means are close to each other, post hoc comparisons using Tamhane's T2 test conducted by SPSS 25.0 (shown in APPENDIX A) indicate that the mean of each category is significantly different ($p < 0.001$) from all other groups. Thus, these distinct features of different transcript groups may be biologically significant.

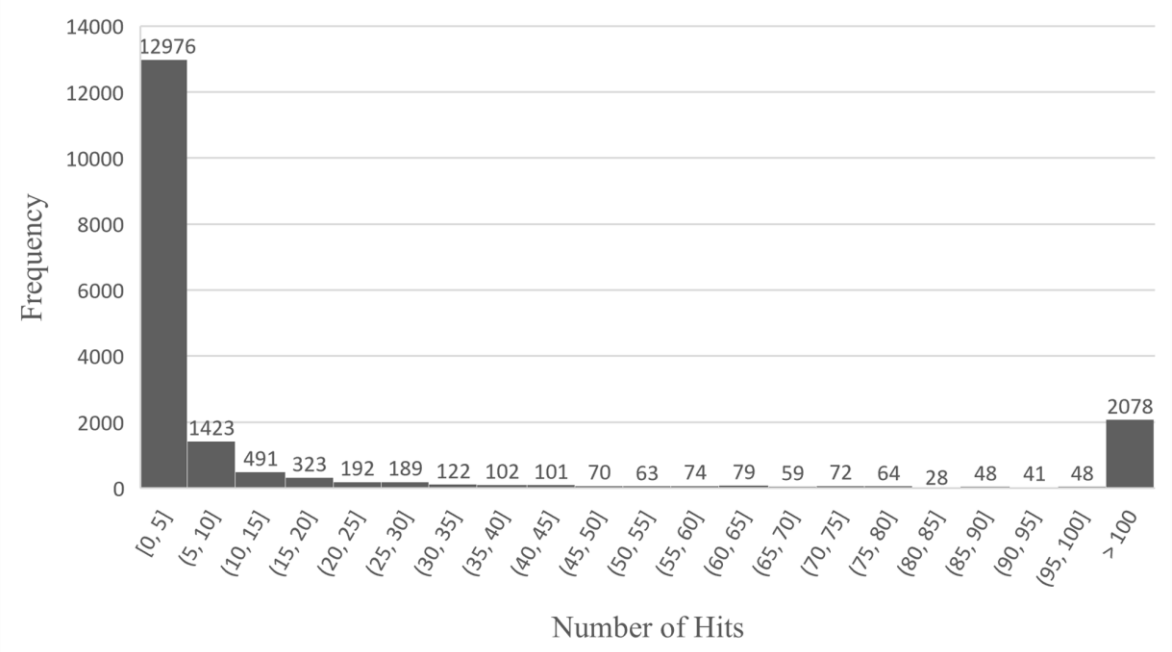

Figure 3.1. Frequency distribution of the number of hits per query. $\bar{x}$ = 237.7, Q50 = 2.0, SD = 2595.9, min = 1.0, max = 153028.0, n = 18,643. $\bar{x}$, Q50, SD, min, max, and n denote mean, median, standard deviation, minimum, maximum, and sample size, respectively.

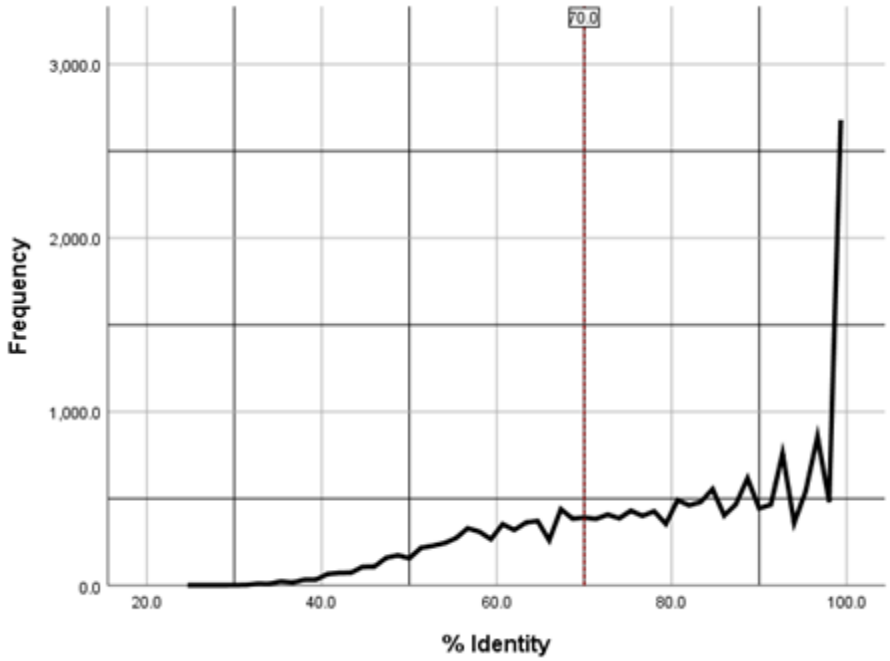

Figure 3.2. Line graph of the top hit % identity of altProts by all RNA types. $\bar{x}$ = 79.4, SD = 16.6, n = 18,643. The bin size was set to one. The X-axis reference line at 70.0 shows the threshold for robust conservation evidence and chimeric proteins and mosaic protein analysis.

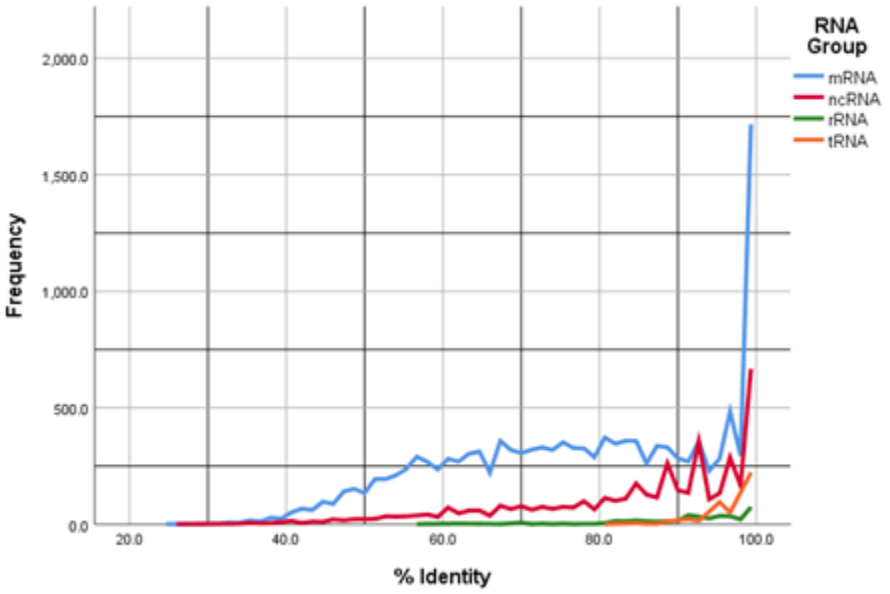

Figure 3.3. Multiple line graph of the top hit % identity of altProts by individual RNA types. mRNA: $\bar{x}$ = 77.0, SD = 16.7, n = 13,427; ncRNA: $\bar{x}$ = 83.8, SD = 14.9, n = 4,398; rRNA: $\bar{x}$ = 91.6, SD = 7.9, n = 392; tRNA: $\bar{x}$ = 97.2, SD = 3.6, n = 426. The bin size was set to one. The X-axis reference line at 70.0 shows the threshold for robust conservation evidence and chimeric protein and mosaic protein analysis.

For robust conservation analysis and detection of chimeric proteins and mosaic proteins, altProts are not considered if their top hit % identities are less than 70%, which eliminates altProts with low similarity (<70% identity) to the reference proteome database. The 70% threshold is shown in Figure 3.2 and Figure 3.3 with a red dash line crossing the X-axis. During filtering, approximately 35% and 20% of mRNA- and ncRNA-derived altProts were eliminated, respectively. However, only 2% of rRNA-derived altProts were eliminated. Interestingly, no tRNA-derived altProts were eliminated as all had % identity values above the threshold. In total, 13,000 altProts remained after eliminating queries with top hits' % identity below 70%. Among them, 8,700 and 3,500 altProts belong to mRNA and ncRNA, respectively, while rRNA and tRNA groups have ~400 altProts per group. Table 3.3 shows exact numbers and percentages of altProts with top hits above or equal to 70% identity (remained) or lower than 70% identity (eliminated).

Table 3.3. Numbers of altProts before and after the elimination of queries with less than 70% identity to annotated proteins

	mRNA	**ncRNA**	**rRNA**	**tRNA**	**Total**
Total altProts with hit	13,427	4,398	392	426	18,643
Eliminated	4,709 (35%)	849 (19%)	7 (2%)	0 (0%)	5,565
Remained	8718 (65%)	3549 (81%)	385 (98%)	426 (100%)	13,078

The 70% identity threshold reduced the number of altProts with 1-5 hits from 13,000 to 8,500 and those with over 100 hits from 2,100 to 1650. This retained 65% of the 1-5 hit group and 79% of the >100 hit group. Overall, altProts associated with the higher number of hits were more likely to pass the 70% identity threshold.

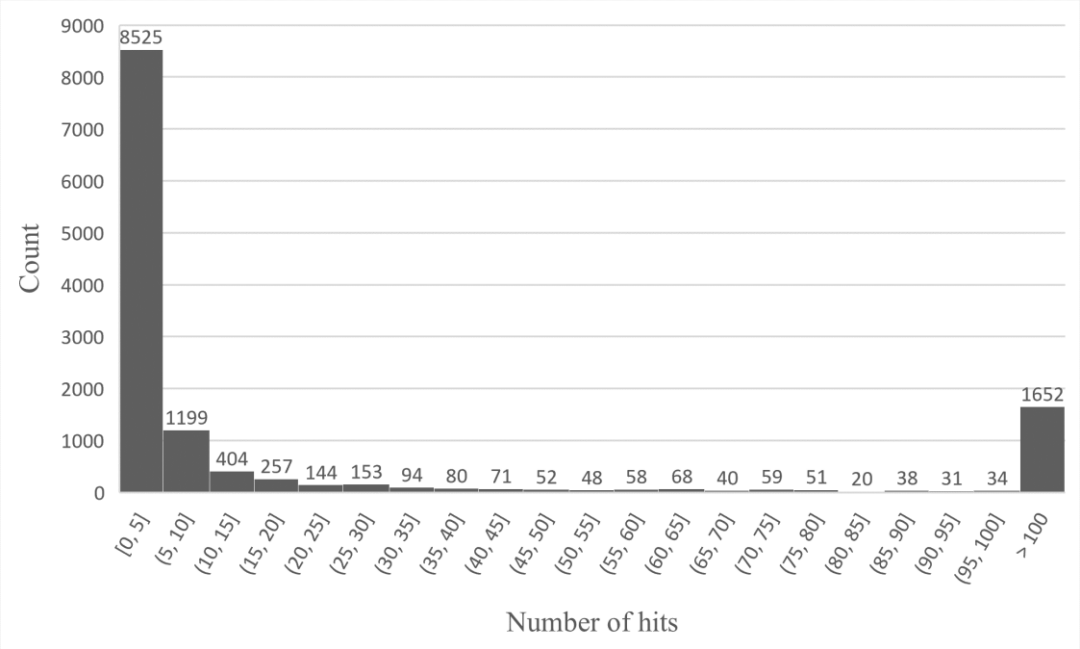

Figure 3.4. Frequency distribution of the number of hits per query after eliminating altProts with the top hit % identity below 70%. $\bar{x}$ =289.5, Q50 = 2.0, SD = 2994.7, min = 1.0, max = 153028.0, n = 13,078.

AltProts with a top hit % identity of 70% or above were candidates for analysis based on conservation evidence and for the identification of chimeric proteins and mosaic proteins. Reference ORFs, also known as canonical ORFs, exist only on mRNA transcripts. Because our software that generates altProts using transcripts as inputs does not discriminate between refORFs and altORFs, protein sequences corresponding to refORFs were determined and removed from the combined primary ORF list. Thus, our candidate altProt list contains no annotated proteins and corresponds only to the currently "unknown" portion of the proteome. AltProts that return hits with at least 70% identity to annotated proteins are conserved compared to the remaining theoretical altProts. This conservation may reflect their functional importance in the model organism *M. truncatula*. Therefore, at least some of these altProts are expected to be translated *in vivo*. These altProts and their corresponding genes are good targets for functional studies on SNF and other fundamental biological processes using loss-of-function methods. The top 100 candidate altProts for each group are available in APPENDIX B. Because the total number of candidate altProts was very large (~13,000), the complete list of candidate altProts will be made available in the public repository after the publication of this data in a peer-reviewed journal; see 2.7 Data Availability.

3.3 Mass Spectrometry-Based Validation of AltProts

To obtain the most direct evidence for translation, all theoretical altProts with a minimum length of 20 aa or longer were subjected to a search using two publicly available MS proteomic datasets with the aid of SearchGUI (v. 4.0.41) and its partner tool Peptide Shaker (v. 2.0.33). MS searches of mRNA, ncRNA, rRNA, and tRNA-derived altProts were performed independently. Because the number of mRNA-derived altProts is too large for regular MS searches, mRNA-derived altProts were searched by the two-step MS approach. In this process, mRNA-derived altProts were split into 10 equal groups, and each group was used independently as a search database. Then, altProts identified from each group were combined. The combined altProts list was used as a search database for the second round of searches on the same dataset. As the number of ncRNA, rRNA, and tRNA-derived altProts did not inflate the search database, the regular one-step procedure was used instead; that is, each type of altProts list (ncRNA, rRNA, and tRNA-derived altProts) was directly searched and validated altProts were recorded for chimeric proteins and mosaic protein identification.

In our protocol, refProts and contaminant databases were included in the search database for MS searches. The rationale behind including refProts and contaminant database was to reveal false positive detections. As expected, most of the validated proteins in each search correspond to refProts. Relatively few validated proteins correspond to altProts. Validated altProts were separated from refProts using their header line; that is, altProts and refProts acquired “altProt” and “refProt” strings in their header line, respectively. When two or more proteins cannot be identified unambiguously by unique peptides, they are grouped in one protein group. If an altProt and a refProt were shown in the same protein group, this altProt is always considered as a non-validated or non-translated altProt group. In other words, an altProt and a refProt may share a common peptide, which is validated by MS searches, and if there is no further validated peptide available to differentiate this altProt and refProt pair, they are grouped in the same protein category. Since refProt are assumed to be translated, altProts grouped with any refProt are not considered as translated altProts.

MS analysis showed that mRNA- and ncRNA-derived altProts were validated; however, rRNA- and tRNA-derived altProts were not validated. The latter two transcript categories were included in these searches because rRNA was previously shown to encode at least six functional polypeptides (Root-Bernstein & Root-Bernstein, 2016). In contrast, tRNA has never been shown to have a protein-coding capacity. However, the presence of highly conserved altProts in the tRNA-derived dataset motivated us to validate these altProts via MS searches. In MS searches for mRNA-derived altProts, 149 (10-day nodules), 98 (14-day nodules), 55 (28-day nodules), 132 (buds), 138 (flowers), 119 (leaves), 73 (roots), 174 (seeds), 96 (stems), and 92 (whole plant) altProts were validated in the first step of the two-step approach. These validated altProts from the first searches were analysed one more time by MS and, in these searches, 125 (10-day nodules), 87 (14-day nodules), 1 (28-day nodules), 119 (buds), 124 (flowers), 100 (leaves), 55 (roots), 138 (seeds), 56 (stems), and 74 (whole plant) altProts were validated. In one-step MS searches for ncRNA-derived altProts, 22 (10-day nodules), 2 (14-day nodules), 0 (28-day nodules), 16 (buds), 12 (flowers), 14 (leaves), 8 (roots), 17 (seeds), 8 (stems), and 11 (whole plant) altProts were validated. The numbers of validated altProts are visualized in Figure 3.5 and also shown in Table 3.4.

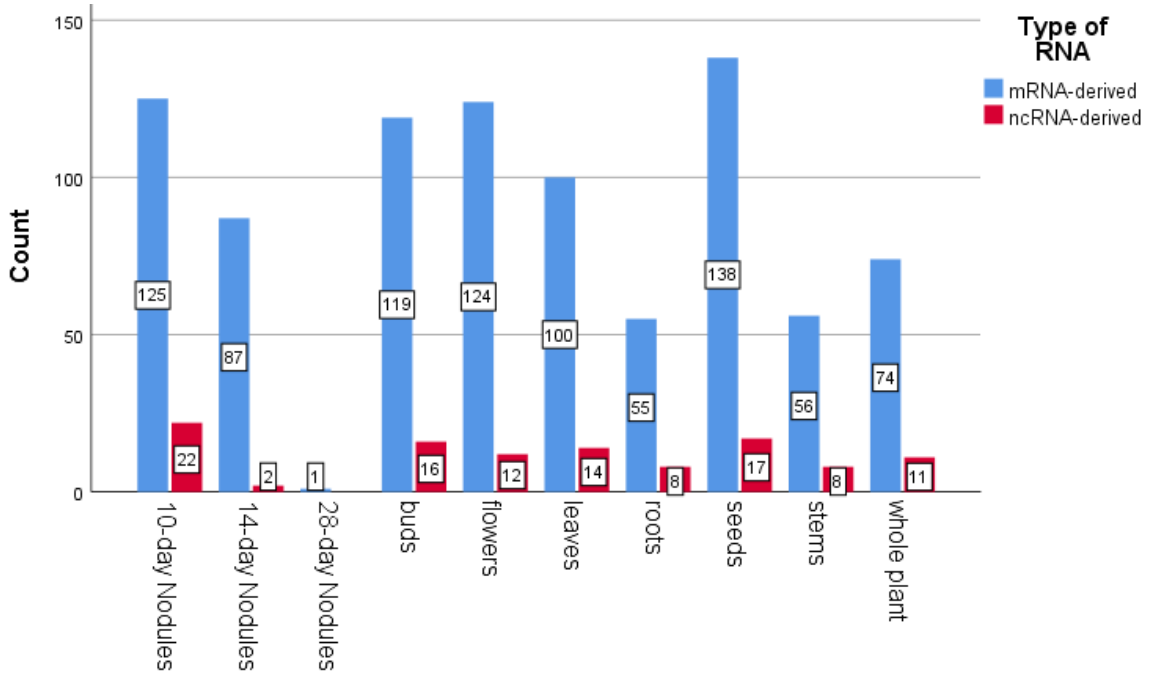

Figure 3.5. Clustered bar counts by type of RNA transcripts for validated altProts in developing nodules and different plant organs. The numbers of validated mRNA-derived altProts are shown from the second search of the two-step procedure, while the MS search for ncRNA-derived altProts consisted of a single step.

Table 3.4. Distribution of validated altProts among various samples

	10-day Nodules	**14-day Nodules**	**28-day Nodules**	**Buds**	**Flowers**	**Leaves**	**Roots**	**Seeds**	**Stems**	**Whole Plant**	**Total Unique** [3]
mRNA-altProts [4]	125 (149)	87 (98)	1 (55)	119 (132)	124 (138)	100 (119)	55 (73)	138 (174)	56 (96)	74 (92)	637
ncRNA-altProts	22	2	0	16	12	14	8	17	8	11	78
PXD002692, Marx et al. (2016)											
PXD013606, Shin et al. (2021)											
Total											715

[3] The number of identified altProts from the second step of the two-step approach was used in the total column.
[4] Numbers in parentheses in the mRNA-altProts row correspond to the validated altProts from the first step of the searching procedure.

While some altProts were validated in only one organ/condition, others were validated in more than one organ/condition. If an altProt was validated only in one organ/condition, it was considered organ/condition-specific. On the other hand, if an altProt was validated in more than one organ/condition or even all cases, it was considered a housekeeping protein translated from alternative open reading frames. 28-day nodules sample has a single validated protein, while younger nodules, especially 10-day nodules, contain many validated proteins. In other organs/conditions, ~100 altProts were detected in total mRNA- and ncRNA-derived altProt analyses.

In total, 715 altProts were validated using two publicly available MS datasets; 637 and 78 altProts were mRNA- and ncRNA-derived altProts, respectively. The list of validated altProts found in more than one organ/condition is shown in APPENDIX C. For the full list, see section 2.7 Data Availability. In MS searches, 513 (mRNA) and 60 (ncRNA) validated altProts were validated in only one organ/condition, and these were considered organ/condition-specific. Then, 61 (mRNA) and seven (ncRNA) altProts were validated in two organs/conditions, and 38 (mRNA) and eight (ncRNA) were validated in three organs/conditions and considered to have housekeeping functions. The number of validated altProts decreased with the increase in the number of organs/conditions in which altProts were validated. Furthermore, 10 (mRNA) and three (ncRNA) altProts were validated in four cases. Then, eight (mRNA) were validated in five cases, but no ncRNA-derived altProt were validated in more than four organs/conditions. No altProt was validated in six organs/conditions; however, six (mRNA) and one (mRNA) altProts were validated in seven and eight cases, respectively. No ncRNA-derived altProts were validated in seven and eight organs/conditions.

Table 3.5. Numbers of altProts that were validated in different organs/conditions.

	Number of organs/conditions in which altProts were validated	**Count**
mRNA	1	513
ncRNA		60
mRNA	2	61
ncRNA		7
mRNA	3	38

Table 3.5. Numbers of altProts that were validated in different organs/conditions. (cont.)

	Number of organs/conditions in which altProts were validated	**Count**
ncRNA		8
mRNA	4	10
ncRNA		3
mRNA	5	8
mRNA	7	6
mRNA	8	1
Total		715

Of the 715 altProts validated by MS, 121 altProts have at least one hit; that is, ~17% of all MS-validated altProts were supported by conservation evidence having 70% similarity to at least one annotated protein. Similar to the whole list of altProts with at least one hit, most MS-supported altProts have either between one and five hits (46) or more than 100 hits (42). The distribution of the number of hits for MS-validated altProts is shown in Figure 3.6. Most of the potential organ/condition-specific altProts (those validated only in one organ/condition) have no hit. However, this does not make them less interesting targets for mutagenesis-based studies as their altORFs may represent *de novo* emerged sequences specific to *M. truncatula*. The potential functional importance of species-specific genes in SNF and other biological processes was recently discussed in the literature (Roy et al., 2020). Out of 573 organ/condition-specific altProts, 56 altProts have at least one hit, which corresponds to ~10%. On the other hand, nearly half (~46%) of the altProts that were validated in more than one organ/condition have hits; that is, 65 out of 142 altProts. Additionally, % identity values of the top hits of MS-supported altProts are mostly very high. For these altProts, the median and the mean values of the top hits are 95.7% and 87.2%, respectively. The heaviest bin in the bar graph shown in Figure 3.7 is 95-100%. It contains 65 MS-supported altProts.

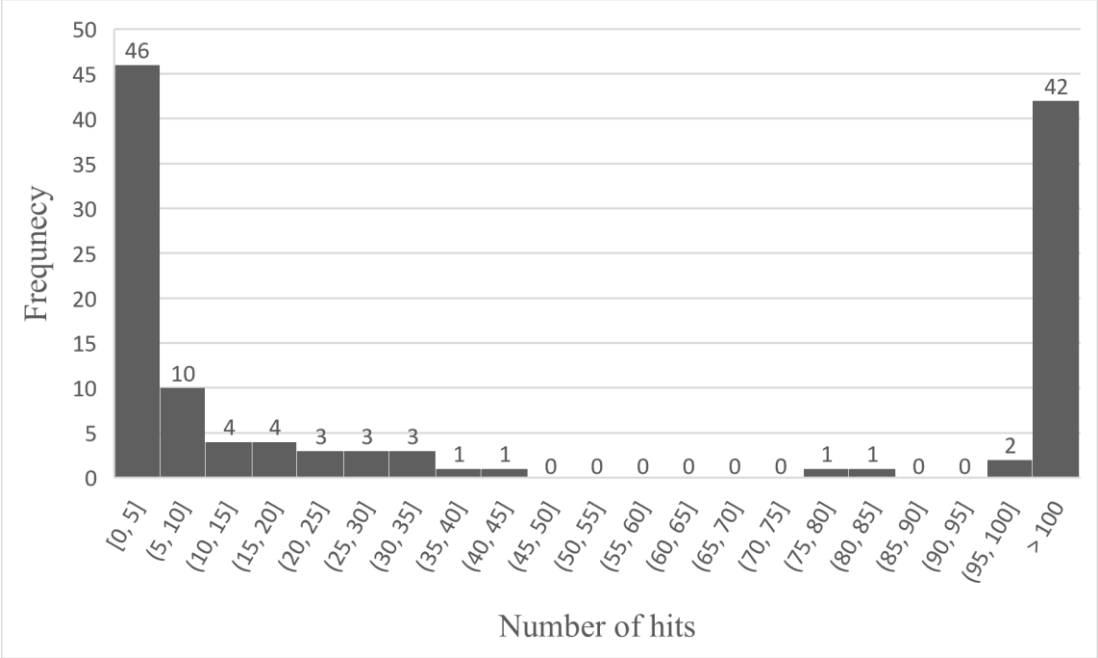

Figure 3.6. Distribution of the number of hits to the reference proteome database per MS-supported altProt. $\bar{x}$ = 2997, Q50 = 18, SD = 17466, n = 121. The bin size was set to five.

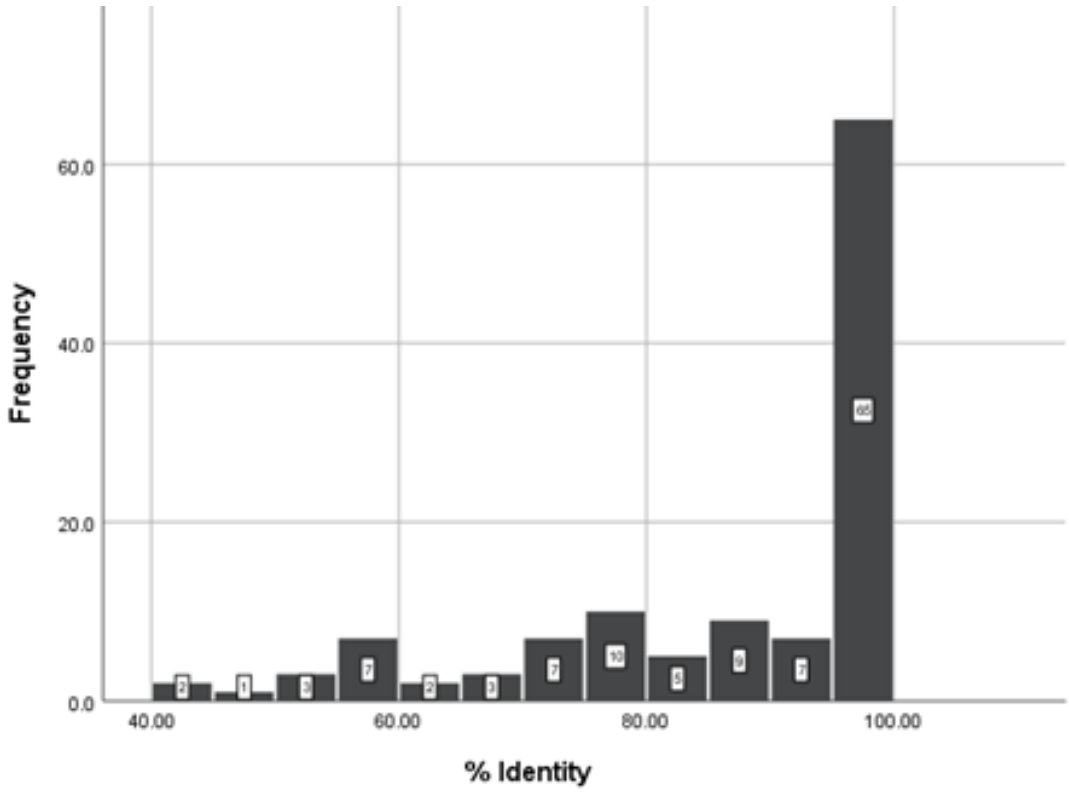

Figure 3.7. Top hit % identity of MS supported altProts to the reference proteome database. $\bar{x}$ = 87.2, Q_{50} = 95.7, SD = 16.0, n = 121. The bin size was set to 5.

3.4 Mass Spectrometry-Based Validation of Chimeric Proteins

Chimeric proteins were modelled from two groups of altProts. The first group was MS-validated altProts that overlap refProts and/or other altProts, and the second group consisted of conserved altProts that overlap refProts and/or other altProts, regardless of the MS validation. For the first group, 715 altProts supported by MS searches were used to model chimeric proteins. Among 715 MS-supported altProts, 636 altProts were derived from mRNA transcripts, and the remaining 78 altProts were derived from ncRNA transcripts. Since every mRNA transcript has one refProt, chimeric proteins were modelled by altProts overlapping with refProts and other altProts. However, ncRNA transcripts do not have a refProt, so, for ncRNA, chimeric proteins were modelled based on altProts that overlap other altProts, and not refProts. For the second group, ~13,100 altProts supported by conservation evidence were used to model chimeric proteins. The top hit % identity of these altProts is at least 70%. For robust analysis, altProts with top hits below 70% identity were excluded. Among altProts that are supported by conservation evidence, ~8,700, ~3500, ~400, and ~400 altProts were derived from mRNA, ncRNA, rRNA, and tRNA transcripts, respectively (see Table 3.3).

Similar to altProt MS searches, refProts and contaminant database were included in the search database of MS searches for chimeric protein validation. Also, MS-validated altProts were included in the search database to avoid false-positive chimeric protein validation. The search parameters used in chimeric protein validation by MS searches were the same as in altProt validation.

Chimeric proteins were modelled according to the following six scenarios depending on whether overlapping takes place between their altORFs and refORFs:

- an altORF is present within its refORF
- an altORF overlaps its refORF at the 5’ end of the refORF
- an altORF overlaps its refORF at the 3’ end of the refORF
- an altORF is located in the 5’UTR of its transcript
- an altORF is located in the 3’UTR of its transcript
- an altORF spans the whole refORF

For a detailed explanation of the modelling of chimeric proteins, please see section 2.4 Modelling of Chimeric Proteins. The six scenarios are also visualized in Figure 2.2.

Chimeric proteins modelled with conserved altProts and MS-validated altProts were validated by MS searches. In total, 147 chimeric proteins were validated, 116 of them belong to chimeric proteins modelled with conserved altProts, and the remaining 31 chimeric proteins belong to chimeric proteins modelled with the MS-validated altProts group. There was one common validated protein (MtrunA17_Chr4g0059001_2F_83-277_195_MtrunA17_ Chr4g0059001_1F_1-840_840) between these two groups. In total, 146 unique chimeric proteins were validated. The distribution of validated chimeric proteins among organs/conditions is shown in Table 3.6.

Table 3.6. Distribution of validated chimeric proteins among various samples

	10-day Nodules	**14-day Nodules**	**28-day Nodules**	**Buds**	**Flowers**	**Leaves**	**Roots**	**Seeds**	**Stems**	**Whole Plant**	**Total Unique** [5]
Chimeric proteins modelled with conserved altProts [6]	22 (48)	16 (24)	0 (2)	16 (26)	19 (26)	20 (39)	11 (15)	32 (52)	12 (31)	74 (20)	116
Chimeric proteins modelled with MS-validated altProts	5	4	0	9	3	6	0	4	1	3	31
PXD002692, Marx et al. (2016)											
PXD013606, Shin et al. (2021)											
Total Unique											146

[5] The number of validated chimeric proteins from the second step of the two-step approach was used in the total column.
[6] Numbers in parentheses show the validated chimeric proteins from the first step of the two-step approach.

3.4.1 Validation of chimeric proteins modelled with MS-validated altProts

MS searches validated 715 altProts, and 715 MS-validated altProts were used to model chimeric proteins. Among MS-validated altProts, 637 altProts were mRNA-derived, and 78 altProts were ncRNA-derived. The location of MS-validated altProt-ORFs relative

to their refProt-ORFs is summarized in Table 3.7. In this table, the first column shows the type of overlapping scenario, as explained above. According to this table, 359 altProt-ORFs (56%) are embedded within their refProt-ORFs (Scenario 1), 49 altProt-ORFs (8%) overlap their refProt-ORFs at the 5' end of the refProt-ORF (Scenario 2), 51 altProt-ORFs (8%) overlap their refProt-ORFs at the 3' end of the refProt-ORF (Scenario 3). Interestingly, while 60 altProt-ORFs (9%) are located in the 5' UTR of their transcripts (Scenario 4), 118 altProt-ORFs (18%) are located in the 3'UTR of their transcripts (Scenario 5).

There is no corresponding refORF for ncRNA-derived altProts. However, the relative position of ncRNA-derived MS-validated altProt-ORFs in their transcripts was determined. For this purpose, each ncRNA sequence was partitioned into three equally sized regions remotely resembling the two UTRs and the refORF of a typical mRNA transcript. Then, the position of each ncRNA-derived altProt-ORF was recorded relative to those three artificial partitions. For instance, an altProt-ORF of type n1 is located in the first one-third of a ncRNA-transcript (n stands for ncRNA); an altProt-ORF of type n12 starts in the first one-third of an ncRNA transcript and ends in the second one-third of its length, and so on. Table 3.7 also shows MS-validated ncRNA-derived altProts categorized according to the location of their ORFs on transcripts. According to this table, 23 altProt-ORFs (30%) are located in the first one-third of the transcript (n1), eight altProt-ORFs (10%) are located in the second one-third of the transcript (n2), and 19 altProt-ORFs (24%) are located in the third one-third of the transcript (n3). Additionally, eight altProt-ORFs (10%) start in the first one-third of the transcript and end in the second one-third of its length (n12), and 14 altProt-ORFs (18%) start in the second one-third of the transcript and end in the third one-third of its length (n23). Furthermore, six altProt-ORFs (8%) start in the first one-third of the transcript and end in the third one-third of its length (n123), and these altProt-ORFs can be considered longer than their "refORFs" (the middle segments).

Table 3.7. ORF positions of MS-validated mRNA and ncRNA-derived altProts relative to their refORFs and the middle portions of their ncRNA transcripts, respectively.

	Scenario	**Count**	**%**
mRNA-derived	1	359	56.4
	2	49	7.7
	3	51	8
	4	60	9.4
	5	118	18.5
	Total	637	100
	Scenario	**Count**	**%**
ncRNA-derived	n1	23	29.5
	n12	8	10.3
	n123	6	7.7
	n2	8	10.3
	n23	14	17.9
	n3	19	24.4
	Total	78	100.0

The algorithm modelled 32,275 chimeric proteins from 715 MS-validated altProts. Although ORFs of seven altProt pairs overlap at the 5’ or 3’ UTR, they could not be used for modelling of chimeric proteins because the gaps between any those altORFs were longer than 10 nt. Thus, all modelled chimeric proteins in this group were altProt-refProt pairs. Among ~32,000 modelled chimeric proteins, 31 chimeric proteins were validated, and their unique identifiers are shown in Table 3.8. In the table, the first column has the row number (RN) to more easily follow the results, and “1” indicates the corresponding chimeric protein was validated and “0” indicates the corresponding chimeric protein was not validated. One chimeric protein (RN1) was validated in three organs/conditions, two chimeric proteins (RN2,3) were validated in two organs/conditions, and the remaining 28 were validated in one organ/condition. Only one chimeric protein inference was as related proteins (at least two MS-validated chimeric proteins have similar or the same sequence that are not differentiated from each other by MS search), and the remaining 30 chimeric proteins were labelled as a single protein. Of note, similar to the altProt MS validation approach, when a protein was labelled as a protein group, only the main accession was taken into consideration, and other accessions in the protein group were not included further. However, other accessions for protein groups are available for interested readers; see section 2.7 Data Availability.

Table 3.8. Validated chimeric proteins.

RN #	Chimeric Proteins	10-day Nodules	14-day Nodules	Buds	Flowers	Leaves	Seeds	Stems	Whole Plant	Total
1	MtrunA17_Chr8g0376411_1F_433-582_150_MtrunA17_Chr8g0376411_3F_156-1463_1308_-1_iteration_6_Within_3'_of_altprot_1505_Chimeric	0	0	1	1	0	0	0	1	3
2	MtrunA17_Chr1g0148371_1F_2233-2337_105_MtrunA17_Chr1g0148371_3F_198-2927_2730_-1_iteration_16_Within_3'_of_altprot_131_Chimeric	1	1	0	0	0	0	0	0	2
3	MtrunA17_Chr1g0162101_3F_3-221_219_MtrunA17_Chr1g0162101_1F_1-255_255_+2_iteration_1_Within_5'_of_altprot_1144_Chimeric	0	0	1	1	0	0	0	0	2
4	MtrunA17_Chr1g0155251_1F_1228-1467_240_MtrunA17_Chr1g0155251_3F_138-2090_1953_-1_iteration_16_Within_3'_of_altprot_691_Chimeric	0	0	0	0	1	0	0	0	1
5	MtrunA17_Chr1g0156271_3F_525-1013_489_MtrunA17_Chr1g0156271_2F_95-3841_3747_-1_iteration_1_Within_3'_of_altprot_760_Chimeric	0	0	0	0	1	0	0	0	1
6	MtrunA17_Chr1g0156271_3F_525-1013_489_MtrunA17_Chr1g0156271_2F_95-3841_3747_+2_iteration_1_Within_3'_of_altprot_781_Chimeric	0	0	1	0	0	0	0	0	1
7	MtrunA17_Chr1g0162101_3F_3-221_219_MtrunA17_Chr1g0162101_1F_1-255_255_+2_iteration_14_Within_5'_of_altprot_1157_Chimeric	0	0	0	0	1	0	0	0	1
8	MtrunA17_Chr1g0178361_1F_529-618_90_MtrunA17_Chr1g0178361_3F_201-1124_924_-1_iteration_9_Within_3'_of_altprot_322_Chimeric	1	0	0	0	0	0	0	0	1
9	MtrunA17_Chr1g0182591_1F_4162-4257_96_MtrunA17_Chr1g0182591_2F_11-4228_4218_-1_iteration_11_3'UTR_overlapped_with_CDS_600_Chimeric	0	0	0	1	0	0	0	0	1
10	MtrunA17_Chr1g0185811_1F_535-621_87_MtrunA17_Chr1g0185811_2F_2-1774_1773_+1_iteration_10_Within_3'_of_altprot_1035_Chimeric	0	0	0	0	0	1	0	0	1
11	MtrunA17_Chr1g0185811_1F_535-621_87_MtrunA17_Chr1g0185811_2F_2-1774_1773_+1_iteration_2_Within_3'_of_altpr	0	0	0	0	0	1	0	0	1

Table 3.8. Validated chimeric proteins. (cont.)

RN #	Chimeric Proteins	10-day Nodules	14-day Nodules	Buds	Flowers	Leaves	Seeds	Stems	Whole Plant	Total
	ot_1027_Chimeric									
12	MtrunA17_Chr1g0185811_1F_535-621_87_MtrunA17_Chr1g0185811_2F_2-1774_1773_+1_iteration_9_Within_3'_of_altprot_1034_Chimeric	0	0	0	0	0	1	0	0	1
13	MtrunA17_Chr2g0298731_1F_3127-3255_129_MtrunA17_Chr2g0298731_3F_105-3674_3570_-1_iteration_7_Within_3'_of_altprot_218_Chimeric	0	0	0	0	0	0	1	0	1
14	MtrunA17_Chr2g0309251_1F_1300-1383_84_MtrunA17_Chr2g0309251_3F_795-2144_1350_-1_iteration_5_Within_3'_of_altprot_476_Chimeric	0	0	1	0	0	0	0	0	1
15	MtrunA17_Chr2g0312631_1F_2020-2094_75_MtrunA17_Chr2g0312631_3F_213-3398_3186_+2_iteration_11_Within_3'_of_altprot_863_Chimeric	0	0	0	0	1	0	0	0	1
16	MtrunA17_Chr3g0083801_1F_76-174_99_MtrunA17_Chr3g0083801_3F_12-737_726_+1_iteration_3_Within_5'_of_altprot_851_Chimeric	1	0	0	0	0	0	0	0	1
17	MtrunA17_Chr3g0091141_1F_976-1071_96_MtrunA17_Chr3g0091141_2F_170-1891_1722_-2_iteration_11_Within_3'_of_altprot_1368_Chimeric	0	0	1	0	0	0	0	0	1
18	MtrunA17_Chr3g0135201_1F_475-633_159_MtrunA17_Chr3g0135201_3F_3-1712_1710_-2_iteration_10_Within_5'_of_altprot_1065_Chimeric	0	0	0	0	0	1	0	0	1
19	MtrunA17_Chr4g0004721_1F_634-813_180_MtrunA17_Chr4g0004721_3F_171-1157_987_+1_iteration_9_Within_5'_of_altprot_248_Chimeric	1	0	0	0	0	0	0	0	1
20	MtrunA17_Chr4g0008511_2F_467-541_75_MtrunA17_Chr4g0008511_1F_1-1500_1500_+2_iteration_14_Within_3'_of_altprot_532_Chimeric	0	0	0	0	1	0	0	0	1
21	MtrunA17_Chr4g0022491_3F_762-923_162_MtrunA17_Chr4g0022491_2F_413-1177_765_+2_iteration_9_Within_3'_of_altprot_889_Chimeric	0	0	0	0	0	0	0	1	1
22	MtrunA17_Chr4g0031721_1F_463-621_159_MtrunA17_Chr4	0	1	0	0	0	0	0	0	1

Table 3.8. Validated chimeric proteins. (cont.)

RN #	Chimeric Proteins	10-day Nodules	14-day Nodules	Buds	Flowers	Leaves	Seeds	Stems	Whole Plant	Total
	g0031721_3F_303-2360_2058_+1_iteration_8_Within_5'_of_altprot_30_Chimeric									
23	MtrunA17_Chr4g0037381_1F_880-1014_135_MtrunA17_Chr4g0037381_2F_422-1024_603_-1_iteration_16_Within_5'_of_altprot_229_Chimeric	0	0	1	0	0	0	0	0	1
24	MtrunA17_Chr4g0045461_2F_215-349_135_MtrunA17_Chr4g0045461_1F_1-1113_1113_+1_iteration_17_Within_5'_of_altprot_943_Chimeric	0	1	0	0	0	0	0	0	1
25	MtrunA17_Chr4g0054921_2F_89-238_150_MtrunA17_Chr4g0054921_1F_1-327_327_-1_iteration_0_Within_3'_of_altprot_107_Chimeric	0	0	0	0	0	0	0	1	1
26	MtrunA17_Chr4g0055331_2F_62-127_66_MtrunA17_Chr4g0055331_1F_1-738_738_+2_iteration_9_Within_3'_of_altprot_269_Chimeric	0	0	1	0	0	0	0	0	1
27	MtrunA17_Chr4g0059001_2F_83-277_195_MtrunA17_Chr4g0059001_1F_1-840_840_-2_iteration_11_Within_5'_of_altprot_576_Chimeric	0	0	1	0	0	0	0	0	1
28	MtrunA17_Chr5g0393951_2F_608-667_60_MtrunA17_Chr5g0393951_1F_301-2001_1701_+2_iteration_3_Within_3'_of_altprot_477_Chimeric	0	1	0	0	0	0	0	0	1
29	MtrunA17_Chr7g0252891_1F_277-459_183_MtrunA17_Chr7g0252891_3F_186-1631_1446_+2_iteration_0_Within_3'_of_altprot_540_Chimeric	0	0	1	0	0	0	0	0	1
30	MtrunA17_Chr7g0259611_2F_185-292_108_MtrunA17_Chr7g0259611_3F_201-779_579_-2_iteration_6_5'UTR_overlapped_with_CDS_795_Chimeric	1	0	0	0	0	0	0	0	1
31	MtrunA17_Chr8g0377071_2F_1406-1693_288_MtrunA17_Chr8g0377071_1F_511-2715_2205_-1_iteration_17_Within_3'_of_altprot_1684_Chimeric	0	0	0	0	1	0	0	0	1

3.4.2 Validation of chimeric proteins modelled with conserved altProts

Chimeric proteins were also generated from the list of conserved altProts; that is, ~13,000 conserved altProts were used to model chimeric proteins. Among ~13,000 conserved altProts, which have at least one hit, ~13,400, ~4400, ~400, and ~400 altProts were generated from mRNA, ncRNA, rRNA, and tRNA groups, respectively (Table 3.3). In total, 533,569 chimeric proteins were generated from the conserved altProts, 324,768 chimeric proteins were altProt-altProt pairs, and the remaining chimeric proteins, 208,801, were altProt-refProt pairs.

ORF positions of conserved altProts relative to ORFs of refProts are shown in Table 3.9. In this table, ~2,700 altProt-ORFs (31%) are embedded within refProt-ORFs (Scenario 1), ~700 altProt-ORFs (8%) overlap their refProt-ORFs at the 5’ end of refProt-ORFs (Scenario 2), ~1,400 altProt-ORFs (16%) overlap their refProt-ORFs at the 3’ end of refProt-ORFs (Scenario 3), ~1,300 altProt-ORFs (14%) are located in the 5’ UTR of their transcripts (Scenario 4), and ~2,700 altProt-ORFs (31%) are located in the 3’ UTR of their transcripts (Scenario 5). Unlike ORFs of MS-validated altProts used for the modelling of chimeric proteins, 44 altProt-ORFs (1%) span their refProt-ORFs (Scenario 6).

In the same table (Table 3.9), the relative positions of non-mRNA-derived conserved altProt-ORFs in their transcripts are shown. As in the case of ncRNA transcripts divided into equal-sized portions n1, n2, and n3, each rRNA and tRNA transcript were partitioned into three equal parts designated r1, r2, r3 and t1, t2, t3, respectively. Among ncRNA-derived conserved altProts, ~820 (23%), ~520 (15%), ~320 (9%), ~650 (18%), ~450 (13%), and ~790 (22%) are members of the n1, n12, n123, n2, n23, and n3 groups, respectively. Among rRNA-derived conserved altProts, 122 (32%), 45 (12%), 16 (4%), 105 (27%), 13 (3%), 83 (22%) are members of the r1, r12, r123, r2, r23, and r3 groups, respectively. Furthermore, all tRNA-derived conserved altProts, ~430, are members of the t123 group, evidently due to the very short average sequence length of tRNA molecules. As in the case of chimeric proteins modelled with MS-validated altProts, a group that includes more than one number indicates that an ORF of the respective altProt spans a border of two segments (e.g., n12) or borders of three segments (e.g., n123).

Table 3.9. ORF positions of conserved altProts relative to their refProt-ORFs (mRNA) or the middle portions of their non-mRNA transcripts (ncRNA, rRNA, and tRNA).

mRNA-derived			ncRNA-derived			rRNA-derived			tRNA-derived		
Scenario	**Count**	**%**	**Scenario**	**Count**	**%**	**Scenario**	**Count**	**%**	**Scenario**	**Count**	**%**
1	2684	30.8	n1	823	23.2	r1	122	31.8			
2	673	7.7	n12	516	14.5	r12	45	11.7			
3	1397	16	n123	318	9	r123	16	4.2	t123	426	100
4	1250	14.3	n2	653	18.4	r2	105	27.3			
5	2670	30.6	n23	447	12.6	r23	13	3.4			
6	44	0.5	n3	791	22.3	r3	83	21.6			
Total	8718	100	Total	3548	100	Total	384	100	Total	426	100

The algorithm modelled ~530,000 chimeric proteins, and the two-step approach was used to search for corresponding MS peptides. Then, 116 chimeric proteins were validated, and their unique identifiers are shown in Table 3.10. Among 116 chimeric proteins, one chimeric protein (RN1) was validated in seven organs/conditions, two chimeric proteins (RN2, RN3) were validated in six organs/conditions, one chimeric protein (RN4) was validated in five organs/conditions, one chimeric protein (RN5) was validated in four organs/conditions, four chimeric proteins (RN6-RN9) were validated in three organs/conditions, 15 chimeric proteins (RN10-RN24) were validated in two organs/conditions, and the remaining 92 chimeric proteins were validated in only one organ/condition, and they were considered as condition-specific.

Table 3.10. Validated chimeric proteins that were modelled by conserved altProts.

RN #	Chimeric Proteins	10-day Nodules	14-day Nodules	Buds	Flowers	Leaves	Roots	Seeds	Stems	Whole Plant	Total
1	MtrunA17_Chr4g0040471_3F_2295-2474_180_MtrunA17_Chr4g0040471_1F_85-2868_2784_-1_iteration_14_Within_5'_of_altprot_6313_Chimeric	1	1	0	1	1	1	1	1	0	7
2	MtrunA17_Chr3g0141901_2F_1316-1585_270_MtrunA17_Chr3g0141901_1F_70-1371_1302_-2_iteration_17_3'UTR_overlapped_with_CDS_6788_Chimeric	1	1	1	1	0	0	1	0	1	6
3	MtrunA17_CPg0492941_3F_825-947_123_MtrunA17_CPg0492941_1F_31-1572_1542_-1_iteration_2_Within_5'_of_altprot_286202_Chimeric	0	1	1	1	1	0	1	0	1	6
4	MtrunA17_MTg0490471_2F_422-538_117_MtrunA17_MTg0490471_1F_220-1740_1521_+1_iteration_16_Within_5'_of_altprot_299661_Chimeric	1	1	0	1	1	0	1	0	0	5

Table 3.10. Validated chimeric proteins that were modelled by conserved altProts. (cont.)

RN #	Chimeric Proteins	10-day Nodules	14-day Nodules	Buds	Flowers	Leaves	Roots	Seeds	Stems	Whole Plant	Total
5	MtrunA17_Chr3g0144151_1F_1222-1311_90_MtrunA17_Chr3g0144151_3F_33-1415_1383_-1_iteration_5_Within_3'_of_altprot_8319_Chimeric	0	0	0	1	1	0	0	1	1	4
6	MtrunA17_Chr3g0105981_2F_1454-1618_165_MtrunA17_Chr3g0105981_3F_27-1478_1452_-1_iteration_5_3'UTR_overlapped_with_CDS_3049_Chimeric	0	1	1	0	0	0	1	0	0	3
7	MtrunA17_Chr3g0144151_1F_1222-1311_90_MtrunA17_Chr3g0144151_3F_33-1415_1383_-1_iteration_6_Within_3'_of_altprot_8320_Chimeric	0	0	0	0	1	0	0	1	1	3
8	MtrunA17_Chr6g0458091_1F_1240-1362_123_MtrunA17_Chr6g0458091_2F_275-1621_1347_-2_iteration_4_Within_3'_of_altprot_7279_Chimeric	1	1	0	0	0	0	1	0	0	3
9	MtrunA17_Chr7g0270811_3F_1536-1718_183_MtrunA17_Chr7g0270811_2F_8-1711_1704_-2_iteration_10_3'UTR_overlapped_with_CDS_937_Chimeric	0	1	1	0	0	0	1	0	0	3
10	MtrunA17_Chr1g0152521_2F_290-436_147_MtrunA17_Chr1g0152521_1F_1-969_969_-2_iteration_20_Within_5'_of_altprot_4039_Chimeric	0	1	0	0	0	0	1	0	0	2
11	MtrunA17_Chr1g0181761_1F_445-570_126_MtrunA17_Chr1g0181761_3F_18-1262_1245_+2_iteration_0_Within_3'_of_altprot_4316_Chimeric	1	0	0	0	0	0	0	0	1	2
12	MtrunA17_Chr1g0202001_2F_149-421_273_MtrunA17_Chr1g0202001_1F_1-2331_2331_-1_iteration_0_Within_3'_of_altprot_2661_Chimeric	1	1	0	0	0	0	0	0	0	2
13	MtrunA17_Chr1g0205601_3F_321-395_75_MtrunA17_Chr1g0205601_1F_13-369_357_-1_iteration_17_3'UTR_overlapped_with_CDS_3876_Chimeric	0	0	0	1	0	1	0	0	0	2
14	MtrunA17_Chr1g0207921_2F_260-391_132_MtrunA17_Chr1g0207921_1F_1-597_597_+1_iteration_6_Within_5'_of_altprot_5369_Chimeric	1	1	0	0	0	0	0	0	0	2
15	MtrunA17_Chr1g0212961_2F_617-703_87_MtrunA17_Chr1g0212961_3F_63-1136_1074_-2_iteration_2_Within_3'_of_altprot_7662_Chimeric	0	0	0	1	1	0	0	0	0	2
16	MtrunA17_Chr4g0070011_3F_1743-1832_90_MtrunA17_Chr4g0070011_1F_58-2481_2424_-1_iteration_16_Within_5'_of_altprot_12584_Chimeric	0	1	0	0	0	1	0	0	0	2
17	MtrunA17_Chr5g0393401_2F_1649-1765_117_MtrunA17_Chr5g0393401_3F_282-1781_1500_-2_iteration_10_Within_3'_of_altprot_2964_Chimeric	1	0	0	1	0	0	0	0	0	2
18	MtrunA17_Chr5g0431401_3F_1149-1388_240_MtrunA17_Chr5g0431401_2F_2-1354_1353_+1_iteration_18_3'UTR_overlapped_with_CDS_6287_Chimeric	0	0	0	0	0	1	0	1	0	2
19	MtrunA17_Chr6g0457461_3F_438-596_159_MtrunA17_Chr6g0457461_2F_170-715_546_-2_iteration_17_Within_5'_of_altprot_6844_Chimeric	0	0	1	0	0	0	1	0	0	2
20	MtrunA17_Chr6g0457461_3F_438-596_159_MtrunA17_Chr6g0457461_2F_170-715_546_-2_iteration_18_Within_5'_of_altprot_6845_Chimeric	0	0	1	0	0	0	0	0	1	2
21	MtrunA17_Chr8g0339891_2F_998-1174_177_MtrunA17_Chr8g0339891_3F_426-1067_642_+2_iteration_11_3'UTR_overlapped_with_CDS_6333_Chimeric	0	0	1	0	0	0	0	0	1	2
22	MtrunA17_Chr8g0345421_3F_2838-2963_126_MtrunA17_Chr8g0345421_2F_83-3610_3528_-2_iteration_6_Within_5'_of_altprot_9266_Chimeric	0	0	0	0	1	0	0	1	0	2
23	MtrunA17_CPg0492331_2F_584-874_291_MtrunA17_CPg0492331_3F_603-704_102_-1_iteration_19_Spanned_3'_of_altprot_278150_Chimeric	0	0	0	0	1	0	0	1	0	2
24	MtrunA17_MTg0491711_1F_2815-2910_96_MtrunA17_MTg0491711_3F_2664-2906_243_+1_iteration_24_3'UTR_overlapped_with_CDS_334975_Chimeric	0	0	0	0	1	0	0	1	0	2
25	MtrunA17_Chr0c01g0489091_1F_199-492_294_MtrunA17_Chr0c01g0489091_	0	0	0	0	1	0	0	0	0	1

Table 3.10. Validated chimeric proteins that were modelled by conserved altProts. (cont.)

RN #	Chimeric Proteins	10-day Nodules	14-day Nodules	Buds	Flowers	Leaves	Roots	Seeds	Stems	Whole Plant	Total
	3F_384-629_246_-1_iteration_2_5'UTR_overlapped_with_CDS_6083_Chimeric										
26	MtrunA17_Chr0c28g0493951_2F_422-661_240_MtrunA17_Chr0c28g0493951_3F_498-1307_810_-2_iteration_6_5'UTR_overlapped_with_CDS_1094_Chimeric	0	0	0	0	0	0	1	0	0	1
27	MtrunA17_Chr1g0149991_2F_1847-2362_516_MtrunA17_Chr1g0149991_3F_171-1910_1740_+2_iteration_0_3'UTR_overlapped_with_CDS_3304_Chimeric	0	0	0	0	0	1	0	0	0	1
28	MtrunA17_Chr1g0150571_3F_1035-1250_216_MtrunA17_Chr1g0150571_2F_224-1192_969_-2_iteration_15_3'UTR_overlapped_with_CDS_3594_Chimeric	0	0	0	0	0	0	1	0	0	1
29	MtrunA17_Chr1g0153001_1F_178-519_342_MtrunA17_Chr1g0153001_3F_396-806_411_-1_iteration_9_5'UTR_overlapped_with_CDS_16607_Chimeric	1	0	0	0	0	0	0	0	0	1
30	MtrunA17_Chr1g0158341_2F_758-859_102_MtrunA17_Chr1g0158341_1F_1-1251_1251_+2_iteration_17_Within_3'_of_altprot_5997_Chimeric	0	0	0	0	0	0	0	0	1	1
31	MtrunA17_Chr1g0164591_3F_3-338_336_MtrunA17_Chr1g0164591_2F_314-1369_1056_-1_iteration_2_5'UTR_overlapped_with_CDS_22564_Chimeric	0	0	0	0	0	1	0	0	0	1
32	MtrunA17_Chr1g0183001_1F_2749-2943_195_MtrunA17_Chr1g0183001_3F_375-4916_4542_-2_iteration_16_Within_5'_of_altprot_29974_Chimeric	1	0	0	0	0	0	0	0	0	1
33	MtrunA17_Chr1g0190571_2F_62-262_201_MtrunA17_Chr1g0190571_1F_1-2274_2274_+2_iteration_10_Within_3'_of_altprot_8324_Chimeric	0	0	0	0	0	0	1	0	0	1
34	MtrunA17_Chr1g0191411_1F_1660-1869_210_MtrunA17_Chr1g0191411_2F_1601-1732_132_+2_iteration_22_3'UTR_overlapped_with_CDS_30714_Chimeric	0	0	0	0	0	0	0	1	0	1
35	MtrunA17_Chr1g0198091_3F_255-443_189_MtrunA17_Chr1g0198091_1F_379-507_129_+1_iteration_2_5'UTR_overlapped_with_CDS_32192_Chimeric	0	0	0	0	0	0	0	0	1	1
36	MtrunA17_Chr1g0200071_3F_828-1001_174_MtrunA17_Chr1g0200071_2F_218-1567_1350_-2_iteration_11_Within_5'_of_altprot_1540_Chimeric	0	0	0	0	0	0	1	0	0	1
37	MtrunA17_Chr1g0200071_3F_828-1001_174_MtrunA17_Chr1g0200071_2F_218-1567_1350_+2_iteration_1_Within_3'_of_altprot_1593_Chimeric	0	0	0	0	1	0	0	0	0	1
38	MtrunA17_Chr1g0207811_1F_1462-1824_363_MtrunA17_Chr1g0207811_3F_207-1913_1707_+1_iteration_18_Within_5'_of_altprot_5233_Chimeric	0	0	0	0	1	0	0	0	0	1
39	MtrunA17_Chr1g0209791_2F_2-130_129_MtrunA17_Chr1g0209791_1F_1-288_288_+2_iteration_5_Within_3'_of_altprot_6420_Chimeric	0	0	0	0	0	0	1	0	0	1
40	MtrunA17_Chr1g0210521_2F_782-928_147_MtrunA17_Chr1g0210521_3F_192-917_726_-1_iteration_23_3'UTR_overlapped_with_CDS_6868_Chimeric	1	0	0	0	0	0	0	0	0	1
41	MtrunA17_Chr2g0283311_1F_766-849_84_MtrunA17_Chr2g0283311_2F_263-868_606_-2_iteration_14_Within_3'_of_altprot_529_Chimeric	0	0	0	0	0	1	0	0	0	1
42	MtrunA17_Chr2g0285461_2F_1553-1795_243_MtrunA17_Chr2g0285461_1F_838-3573_2736_+2_iteration_0_Within_3'_of_altprot_1397_Chimeric	0	1	0	0	0	0	0	0	0	1
43	MtrunA17_Chr2g0292921_2F_1202-1453_252_MtrunA17_Chr2g0292921_1F_187-1545_1359_-2_iteration_6_Within_5'_of_altprot_5049_Chimeric	0	0	0	0	0	0	1	0	0	1
44	MtrunA17_Chr2g0299561_2F_3347-3535_189_MtrunA17_Chr2g0299561_1F_97-3969_3873_-1_iteration_17_Within_3'_of_altprot_7469_Chimeric	1	0	0	0	0	0	0	0	0	1
45	MtrunA17_Chr2g0304891_3F_627-725_99_MtrunA17_Chr2g0304891_1F_643-	0	0	0	0	0	0	0	0	1	1

Table 3.10. Validated chimeric proteins that were modelled by conserved altProts. (cont.)

RN #	Chimeric Proteins	10-day Nodules	14-day Nodules	Buds	Flowers	Leaves	Roots	Seeds	Stems	Whole Plant	Total
	1512_870_-2_iteration_1_5'UTR_overlapped_with_CDS_47848_Chimeric										
46	MtrunA17_Chr2g0305951_3F_3-113_111_MtrunA17_Chr2g0305951_1F_1-153_153_-1_iteration_14_Within_5'_of_altprot_10021_Chimeric	0	1	0	0	0	0	0	0	0	1
47	MtrunA17_Chr2g0316291_3F_1647-1745_99_MtrunA17_Chr2g0316291_2F_101-3394_3294_-2_iteration_1_Within_5'_of_altprot_2908_Chimeric	0	0	0	0	0	0	0	0	1	1
48	MtrunA17_Chr2g0326801_1F_2404-2493_90_MtrunA17_Chr2g0326801_3F_2394-2507_114_-2_iteration_19_Within_5'_of_altprot_62556_Chimeric	1	0	0	0	0	0	0	0	0	1
49	MtrunA17_Chr2g0328091_3F_87-245_159_MtrunA17_Chr2g0328091_2F_107-1243_1137_-1_iteration_5_5'UTR_overlapped_with_CDS_8150_Chimeric	0	0	0	1	0	0	0	0	0	1
50	MtrunA17_Chr2g0329031_3F_1209-1379_171_MtrunA17_Chr2g0329031_2F_197-2440_2244_+2_iteration_20_Within_3'_of_altprot_8576_Chimeric	0	0	1	0	0	0	0	0	0	1
51	MtrunA17_Chr3g0096421_3F_54-173_120_MtrunA17_Chr3g0096421_1F_46-1086_1041_-1_iteration_7_Within_5'_of_altprot_8570_Chimeric	0	0	0	0	0	0	1	0	0	1
52	MtrunA17_Chr3g0100221_2F_803-922_120_MtrunA17_Chr3g0100221_1F_34-852_819_+1_iteration_6_3'UTR_overlapped_with_CDS_638_Chimeric	0	0	0	0	0	1	0	0	0	1
53	MtrunA17_Chr3g0102171_2F_236-379_144_MtrunA17_Chr3g0102171_1F_73-3423_3351_-2_iteration_5_Within_5'_of_altprot_1400_Chimeric	0	0	0	0	0	0	1	0	0	1
54	MtrunA17_Chr3g0113591_2F_1295-1465_171_MtrunA17_Chr3g0113591_1F_124-2748_2625_-2_iteration_19_Within_5'_of_altprot_5397_Chimeric	0	0	0	0	1	0	0	0	0	1
55	MtrunA17_Chr3g0124631_2F_1127-1249_123_MtrunA17_Chr3g0124631_1F_130-2304_2175_+2_iteration_19_Within_3'_of_altprot_9937_Chimeric	0	0	0	1	0	0	0	0	0	1
56	MtrunA17_Chr3g0130671_3F_1026-1190_165_MtrunA17_Chr3g0130671_1F_334-1179_846_-1_iteration_5_3'UTR_overlapped_with_CDS_1358_Chimeric	1	0	0	0	0	0	0	0	0	1
57	MtrunA17_Chr3g0135761_2F_734-958_225_MtrunA17_Chr3g0135761_1F_1-2115_2115_+2_iteration_0_Within_3'_of_altprot_3681_Chimeric	0	0	0	0	0	0	1	0	0	1
58	MtrunA17_Chr3g0137391_3F_915-1070_156_MtrunA17_Chr3g0137391_2F_23-1768_1746_-1_iteration_8_Within_3'_of_altprot_92040_Chimeric	0	0	1	0	0	0	0	0	0	1
59	MtrunA17_Chr3g0144151_1F_1222-1311_90_MtrunA17_Chr3g0144151_3F_33-1415_1383_+2_iteration_3_Within_3'_of_altprot_8338_Chimeric	0	0	0	0	1	0	0	0	0	1
60	MtrunA17_Chr4g0000131_1F_1471-1644_174_MtrunA17_Chr4g0000131_3F_159-2459_2301_+1_iteration_7_Within_5'_of_altprot_9208_Chimeric	1	0	0	0	0	0	0	0	0	1
61	MtrunA17_Chr4g0000891_3F_3-473_471_MtrunA17_Chr4g0000891_1F_1-471_471_+2_iteration_6_3'UTR_overlapped_with_CDS_9994_Chimeric	0	1	0	0	0	0	0	0	0	1
62	MtrunA17_Chr4g0014251_3F_45-242_198_MtrunA17_Chr4g0014251_1F_1-327_327_+1_iteration_4_Within_3'_of_altprot_7125_Chimeric	1	0	0	0	0	0	0	0	0	1
63	MtrunA17_Chr4g0016371_3F_798-1025_228_MtrunA17_Chr4g0016371_1F_1-1353_1353_-2_iteration_13_Within_3'_of_altprot_118439_Chimeric	0	0	0	0	0	0	0	1	0	1
64	MtrunA17_Chr4g0016371_3F_798-1025_228_MtrunA17_Chr4g0016371_1F_1-1353_1353_-2_iteration_20_Within_3'_of_altprot_118446_Chimeric	0	0	0	0	0	0	0	1	0	1
65	MtrunA17_Chr4g0023791_1F_1-84_84_MtrunA17_Chr4g0023791_2F_2-82_81_-1_iteration_5_Spanned_3'_of_altprot_126958_Chimeric	0	0	0	1	0	0	0	0	0	1

Table 3.10. Validated chimeric proteins that were modelled by conserved altProts. (cont.)

RN #	Chimeric Proteins	10-day Nodules	14-day Nodules	Buds	Flowers	Leaves	Roots	Seeds	Stems	Whole Plant	Total
66	MtrunA17_Chr4g0034491_2F_1151-1708_558_MtrunA17_Chr4g0034491_1F_88-2415_2328_-2_iteration_1_Within_5'_of_altprot_4430_Chimeric	0	0	0	1	0	0	0	0	0	1
67	MtrunA17_Chr4g0059001_2F_83-277_195_MtrunA17_Chr4g0059001_1F_1-840_840_-2_iteration_11_Within_5'_of_altprot_6822_Chimeric	0	0	1	0	0	0	0	0	0	1
68	MtrunA17_Chr4g0063201_2F_446-535_90_MtrunA17_Chr4g0063201_1F_262-522_261_-2_iteration_18_3'UTR_overlapped_with_CDS_8999_Chimeric	0	0	0	0	0	0	1	0	0	1
69	MtrunA17_Chr5g0405061_2F_182-412_231_MtrunA17_Chr5g0405061_3F_228-2174_1947_-2_iteration_20_5'UTR_overlapped_with_CDS_7846_Chimeric	1	0	0	0	0	0	0	0	0	1
70	MtrunA17_Chr5g0415031_1F_931-1152_222_MtrunA17_Chr5g0415031_3F_114-4400_4287_+1_iteration_19_Within_5'_of_altprot_1210_Chimeric	0	0	0	0	1	0	0	0	0	1
71	MtrunA17_Chr5g0415911_1F_3553-3948_396_MtrunA17_Chr5g0415911_3F_2832-3632_801_-2_iteration_6_3'UTR_overlapped_with_CDS_152972_Chimeric	1	0	0	0	0	0	0	0	0	1
72	MtrunA17_Chr5g0421761_2F_1835-1918_84_MtrunA17_Chr5g0421761_3F_1818-1925_108_-1_iteration_4_Within_5'_of_altprot_158546_Chimeric	0	0	0	0	0	0	1	0	0	1
73	MtrunA17_Chr5g0422291_2F_3056-3259_204_MtrunA17_Chr5g0422291_3F_3120-3197_78_-1_iteration_5_Spanned_3'_of_altprot_177338_Chimeric	0	0	1	0	0	0	0	0	0	1
74	MtrunA17_Chr5g0430341_2F_2501-2623_123_MtrunA17_Chr5g0430341_3F_2298-2531_234_+2_iteration_3_3'UTR_overlapped_with_CDS_188915_Chimeric	0	0	0	0	1	0	0	0	0	1
75	MtrunA17_Chr5g0430341_3F_1059-1172_114_MtrunA17_Chr5g0430341_1F_1-1620_1620_-2_iteration_8_Within_3'_of_altprot_188974_Chimeric	0	0	0	0	0	0	0	0	1	1
76	MtrunA17_Chr5g0435191_1F_883-1032_150_MtrunA17_Chr5g0435191_2F_677-979_303_-1_iteration_25_3'UTR_overlapped_with_CDS_190644_Chimeric	0	0	0	0	0	0	1	0	0	1
77	MtrunA17_Chr5g0444231_2F_758-967_210_MtrunA17_Chr5g0444231_3F_3-902_900_+2_iteration_11_3'UTR_overlapped_with_CDS_195106_Chimeric	0	0	0	1	0	0	0	0	0	1
78	MtrunA17_Chr6g0451601_2F_434-577_144_MtrunA17_Chr6g0451601_1F_91-543_453_-2_iteration_29_3'UTR_overlapped_with_CDS_4792_Chimeric	0	0	0	1	0	0	0	0	0	1
79	MtrunA17_Chr6g0452781_3F_4245-4397_153_MtrunA17_Chr6g0452781_1F_1-4458_4458_-1_iteration_15_Within_5'_of_altprot_5038_Chimeric	0	0	0	1	0	0	0	0	0	1
80	MtrunA17_Chr6g0457461_3F_438-596_159_MtrunA17_Chr6g0457461_2F_170-715_546_+2_iteration_6_Within_3'_of_altprot_6896_Chimeric	0	0	1	0	0	0	0	0	0	1
81	MtrunA17_Chr6g0459481_3F_183-263_81_MtrunA17_Chr6g0459481_1F_1-363_363_-1_iteration_12_Within_5'_of_altprot_7678_Chimeric	0	0	0	0	1	0	0	0	0	1
82	MtrunA17_Chr6g0462271_3F_3-149_147_MtrunA17_Chr6g0462271_1F_1-180_180_+1_iteration_5_Within_3'_of_altprot_8508_Chimeric	0	0	0	0	0	0	0	0	1	1
83	MtrunA17_Chr6g0468411_3F_408-509_102_MtrunA17_Chr6g0468411_1F_1-507_507_+2_iteration_28_3'UTR_overlapped_with_CDS_1284_Chimeric	0	0	0	0	0	0	1	0	0	1
84	MtrunA17_Chr6g0476751_3F_2274-2426_153_MtrunA17_Chr6g0476751_2F_359-2653_2295_+2_iteration_5_Within_3'_of_altprot_4139_Chimeric	0	0	0	0	0	1	0	0	0	1
85	MtrunA17_Chr6g0485321_1F_1768-1881_114_MtrunA17_Chr6g0485321_3F_1674-1892_219_+2_iteration_13_Within_3'_of_altprot_225709_Chimeric	0	0	0	0	0	1	0	0	0	1
86	MtrunA17_Chr6g0486961_3F_1875-2120_246_MtrunA17_Chr6g0486961_1F_1	0	0	0	0	0	0	1	0	0	1

Table 3.10. Validated chimeric proteins that were modelled by conserved altProts. (cont.)

RN #	Chimeric Proteins	10-day Nodules	14-day Nodules	Buds	Flowers	Leaves	Roots	Seeds	Stems	Whole Plant	Total
	-2943_2943_-2_iteration_12_Within_3'_of_altprot_8857_Chimeric										
87	MtrunA17_Chr7g0214741_1F_340-462_123_MtrunA17_Chr7g0214741_3F_132-1478_1347_+1_iteration_8_Within_5'_of_altprot_228525_Chimeric	0	0	0	0	0	0	1	0	0	1
88	MtrunA17_Chr7g0214911_3F_183-368_186_MtrunA17_Chr7g0214911_2F_140-1999_1860_-2_iteration_6_Within_5'_of_altprot_228670_Chimeric	0	0	0	0	0	0	1	0	0	1
89	MtrunA17_Chr7g0219691_1F_1507-1602_96_MtrunA17_Chr7g0219691_3F_12-2699_2688_-2_iteration_2_Within_5'_of_altprot_2037_Chimeric	0	0	0	1	0	0	0	0	0	1
90	MtrunA17_Chr7g0221631_3F_3-125_123_MtrunA17_Chr7g0221631_1F_1-351_351_-1_iteration_13_Within_5'_of_altprot_2508_Chimeric	0	0	1	0	0	0	0	0	0	1
91	MtrunA17_Chr7g0230341_2F_2-307_306_MtrunA17_Chr7g0230341_1F_16-186_171_-1_iteration_3_Spanned_5'_of_altprot_5545_Chimeric	1	0	0	0	0	0	0	0	0	1
92	MtrunA17_Chr7g0232851_1F_502-684_183_MtrunA17_Chr7g0232851_2F_587-775_189_-2_iteration_22_5'UTR_overlapped_with_CDS_6181_Chimeric	1	0	0	0	0	0	0	0	0	1
93	MtrunA17_Chr7g0251971_3F_3-161_159_MtrunA17_Chr7g0251971_1F_1-189_189_-1_iteration_4_Within_5'_of_altprot_4082_Chimeric	0	0	0	0	0	0	0	1	0	1
94	MtrunA17_Chr7g0262821_3F_1662-1847_186_MtrunA17_Chr7g0262821_2F_86-2131_2046_-1_iteration_13_Within_3'_of_altprot_8551_Chimeric	0	0	0	0	0	0	1	0	0	1
95	MtrunA17_Chr7g1034306_1F_502-681_180_MtrunA17_Chr7g1034306_2F_533-703_171_-2_iteration_17_5'UTR_overlapped_with_CDS_255945_Chimeric	0	0	0	0	0	0	1	0	0	1
96	MtrunA17_Chr8g0338111_1F_436-534_99_MtrunA17_Chr8g0338111_2F_476-733_258_-2_iteration_11_5'UTR_overlapped_with_CDS_257459_Chimeric	0	0	0	0	0	0	1	0	0	1
97	MtrunA17_Chr8g0338301_2F_1574-1681_108_MtrunA17_Chr8g0338301_1F_1-1680_1680_+1_iteration_17_3'UTR_overlapped_with_CDS_5713_Chimeric	0	0	0	0	0	0	1	0	0	1
98	MtrunA17_Chr8g0353711_2F_110-283_174_MtrunA17_Chr8g0353711_1F_1-282_282_+1_iteration_21_3'UTR_overlapped_with_CDS_261472_Chimeric	0	0	0	0	1	0	0	0	0	1
99	MtrunA17_Chr8g0355501_1F_1552-1761_210_MtrunA17_Chr8g0355501_3F_63-1898_1836_-2_iteration_6_Within_5'_of_altprot_3803_Chimeric	0	1	0	0	0	0	0	0	0	1
100	MtrunA17_Chr8g0356581_1F_1387-1464_78_MtrunA17_Chr8g0356581_2F_1394-1516_123_+1_iteration_5_5'UTR_overlapped_with_CDS_261959_Chimeric	0	0	0	0	0	1	0	0	0	1
101	MtrunA17_Chr8g0365341_2F_773-1216_444_MtrunA17_Chr8g0365341_1F_169-2331_2163_+2_iteration_7_Within_3'_of_altprot_7837_Chimeric	0	0	1	0	0	0	0	0	0	1
102	MtrunA17_Chr8g0368931_1F_889-1059_171_MtrunA17_Chr8g0368931_3F_3-1043_1041_-2_iteration_12_3'UTR_overlapped_with_CDS_8971_Chimeric	0	0	0	0	0	0	1	0	0	1
103	MtrunA17_Chr8g0371281_3F_696-851_156_MtrunA17_Chr8g0371281_2F_122-955_834_-2_iteration_4_Within_5'_of_altprot_9944_Chimeric	1	0	0	0	0	0	0	0	0	1
104	MtrunA17_Chr8g0371741_1F_691-888_198_MtrunA17_Chr8g0371741_3F_189-794_606_+1_iteration_16_3'UTR_overlapped_with_CDS_268066_Chimeric	0	0	0	0	1	0	0	0	0	1
105	MtrunA17_Chr8g0373091_1F_457-702_246_MtrunA17_Chr8g0373091_3F_234-1106_873_+1_iteration_3_Within_5'_of_altprot_10876_Chimeric	0	0	0	1	0	0	0	0	0	1
106	MtrunA17_Chr8g0385331_1F_1456-1719_264_MtrunA17_Chr8g0385331_3F_255-1670_1416_-2_iteration_8_3'UTR_overlapped_with_CDS_4525_Chimeric	0	0	1	0	0	0	0	0	0	1

Table 3.10. Validated chimeric proteins that were modelled by conserved altProts. (cont.)

RN #	Chimeric Proteins	10-day Nodules	14-day Nodules	Buds	Flowers	Leaves	Roots	Seeds	Stems	Whole Plant	Total
107	MtrunA17_CPg0492381_1F_631-750_120_MtrunA17_CPg0492381_2F_671-820_150_+1_iteration_1_5'UTR_overlapped_with_CDS_281060_Chimeric	0	1	0	0	0	0	0	0	0	1
108	MtrunA17_CPg0492461_1F_3232-3426_195_MtrunA17_CPg0492461_3F_1413-7163_5751_+1_iteration_12_Within_5'_of_altprot_283943_Chimeric	0	0	0	1	0	0	0	0	0	1
109	MtrunA17_CPg0492851_2F_209-328_120_MtrunA17_CPg0492851_3F_129-1157_1029_+1_iteration_6_Within_3'_of_altprot_285861_Chimeric	0	0	0	1	0	0	0	0	0	1
110	MtrunA17_CPg0493401_2F_1223-1450_228_MtrunA17_CPg0493401_3F_261-1778_1518_-2_iteration_7_Within_3'_of_altprot_289445_Chimeric	0	0	0	0	1	0	0	0	0	1
111	MtrunA17_MTg0490471_2F_1031-1147_117_MtrunA17_MTg0490471_1F_220-1740_1521_+1_iteration_7_Within_5'_of_altprot_299568_Chimeric	0	0	0	0	0	0	1	0	0	1
112	MtrunA17_MTg0490971_2F_1562-1699_138_MtrunA17_MTg0490971_3F_1542-1727_186_+1_iteration_12_Within_3'_of_altprot_311571_Chimeric	0	0	0	0	0	0	0	1	0	1
113	MtrunA17_MTg0490971_2F_476-628_153_MtrunA17_MTg0490971_1F_277-534_258_+1_iteration_0_3'UTR_overlapped_with_CDS_312023_Chimeric	0	0	0	0	0	0	0	0	1	1
114	MtrunA17_MTg0491151_1F_4534-4659_126_MtrunA17_MTg0491151_2F_4511-4612_102_+2_iteration_15_3'UTR_overlapped_with_CDS_315888_Chimeric	0	0	0	0	0	0	1	0	0	1
115	MtrunA17_MTg0491291_2F_1358-1447_90_MtrunA17_MTg0491291_3F_1329-1421_93_-1_iteration_18_3'UTR_overlapped_with_CDS_322145_Chimeric	0	0	1	0	0	0	0	0	0	1
116	MtrunA17_MTg0491621_1F_1660-1836_177_MtrunA17_MTg0491621_3F_1377-1772_396_+1_iteration_16_3'UTR_overlapped_with_CDS_333675_Chimeric	1	0	0	0	0	0	0	0	0	1

3.5 Transcripts Possibly Associated with Mosaic Proteins

Mosaic proteins are produced by more than one ribosomal frameshifting event during the translation on the same transcript. Thus, a mosaic protein must be composed of at least two chimeric proteins. To identify candidates for mosaic proteins, we checked if any transcript is associated with more than one chimeric protein. This search was based on chimeric proteins modelled by both altProt groups: MS-validated altProts and conserved altProts.

3.5.1 Candidate mosaic proteins deduced from chimeric proteins modelled with MS-validated altProts

In this group of chimeric proteins, the modelling was based on MS-validated altProts. Because of the small database size, they were validated by a regular, one-step MS search. In total, 31 chimeric proteins were validated. Then, three transcripts were found to be associated with more than one chimeric protein, as shown in Table 3.11. Two of them are associated with two chimeric proteins per transcript, and one of them is associated with three chimeric proteins per transcript. All altProt-ORFs from these three transcripts are embedded in their refProt-ORFs. The following transcripts are associated with two chimeric proteins each: MtrunA17_Chr1g0162101 and MtrunA17_Chr1g0156271. Transcript MtrunA17_Chr1g0185811 is associated with three chimeric proteins.

The ORF of altProt MtrunA17_Chr1g0162101_3F_3-221_219 is located at the very beginning of its refORF, which corresponds to refProt MtrunA17_Chr1g0162101_1F_1-255_255. As reflected in unique identifiers of these proteins, the refORF starts at position 1 of the transcript and the altORF starts at position 3. The validated chimeric protein starts with only one aa from the refProt at the 5’ end (the first one) and continues with the remaining 73 aa (which is 219/3) from the altProt. If this chimeric protein is translated, the +2 ribosomal frameshift occurs after translating only one aa from the refProt, which corresponds to iteration 1 of the modelling algorithm. Similar to generating the CopA(Z) protein (Meydan et al., 2017), this validated chimeric protein starts with only one aa from one frame. Besides, the same ribosomal frameshift occurs from the third frame to the first frame according to both chimeric proteins denoted with iteration 1 or 14 and no additional chimeric proteins are detected between these two iterations. Thus, since ribosomal frameshift has to change the frame, it cannot be evidence for the mosaic protein expression. However, chimeric protein denoted with iteration 14 modelled by MtrunA17_Chr1g0162101 can be a good candidate for single ribosomal frameshift and evidence for chimeric protein translation.

However, transcripts MtrunA17_Chr1g0156271 and MtrunA17_Chr1g0185811 that are associated with more than one chimeric protein each have ribosomal frameshift positions that are too close to each other. Frameshifting positions were illustrated in

APPENDIX F. In the MtrunA17_Chr1g0156271 transcript, one ribosomal frameshift is -1 (the ribosome slips back) and the other ribosomal frameshift is +2 (the ribosome slips forward). Two chimeric proteins associated with the MtrunA17_Chr1g0156271 transcript are different by only one aa; actually, there is a gap. Even though two chimeric proteins are associated with the MtrunA17_Chr1g0156271 transcript, this transcript could not be considered as evidence for mosaic translation. Similarly, three chimeric proteins are associated with the MtrunA17_Chr1g0185811 transcript, and two of them, iteration 9 and iteration 10, are different by one aa; there is a different aa in the same position. Furthermore, even though two chimeric proteins (iteration 2 and iteration 9 or 10) may seem to be evidence for mosaic translation, ribosomal frameshift occurs on the same frame; that is, ribosomal frameshift occurs from the first frame to the second frame according to the first chimeric protein denoted with iteration 2 associated with MtrunA17_Chr1g0185811. However, another ribosomal frameshift occurs from the first frame to the second frame according to the second chimeric protein denoted with iteration 9 or 10. Unfortunately, no chimeric protein was detected for the ribosomal frameshift from the second frame to the first frame. Similar to the chimeric proteins modelled with the MtrunA17_Chr1g0162101 transcript, since ribosomal frameshift changes the frame, without further evidence, these three chimeric proteins could not show the existence of mosaic proteins. On the other hand, if a suitable ribosomal frameshift between iteration 2 and 9 or 10 is detected, this transcript can be evidence for the mosaic translation hypothesis. Besides, it is still a significant discovery even though these chimeric proteins are not combined in a continuous mosaic protein because three ribosomal frameshifting events per transcript were detected.

Table 3.11. Transcripts that are associated with more than one chimeric protein modelled with MS-validated altProts.

Transcript IDs	**Chimeric Protein IDs**
MtrunA17_Chr1g0162101	(1) MtrunA17_Chr1g0162101_3F_3-221_219_MtrunA17_Chr1g0162101_1F_1-255_255_+2_iteration_1_Within_5'_of_altprot_1144_Chimeric (2) MtrunA17_Chr1g0162101_3F_3-221_219_MtrunA17_Chr1g0162101_1F_1-255_255_+2_iteration_14_Within_5'_of_altprot_1157_Chimeric

Table 3.11. Transcripts that are associated with more than one chimeric protein modelled with MS-validated altProts. (cont.)

Transcript IDs	Chimeric Protein IDs
MtrunA17_Chr1g0156271	(1) MtrunA17_Chr1g0156271_3F_525-1013_489_MtrunA17_Chr1g0156271_2F_95-3841_3747_-1_iteration_1_Within_3'_of_altprot_760_Chimeric (2) MtrunA17_Chr1g0156271_3F_525-1013_489_MtrunA17_Chr1g0156271_2F_95-3841_3747_+2_iteration_1_Within_3'_of_altprot_781_Chimeric
MtrunA17_Chr1g0185811	(1) MtrunA17_Chr1g0185811_1F_535-621_87_MtrunA17_Chr1g0185811_2F_2-1774_1773_+1_iteration_2_Within_3'_of_altprot_1027_Chimeric (2) MtrunA17_Chr1g0185811_1F_535-621_87_MtrunA17_Chr1g0185811_2F_2-1774_1773_+1_iteration_9_Within_3'_of_altprot_1034_Chimeric (3) MtrunA17_Chr1g0185811_1F_535-621_87_MtrunA17_Chr1g0185811_2F_2-1774_1773_+1_iteration_10_Within_3'_of_altprot_1035_Chimeric

3.5.2 Candidate mosaic proteins deduced from chimeric proteins modelled with conserved altProts

In this group of chimeric proteins, the modelling was based on conserved altProts. Because of the large database size, the validation was conducted by the two-step MS search approach. In total, 116 chimeric proteins were validated. Seven transcripts were found to be associated with more than one chimeric protein, as shown in Table 3.12. Five of them are associated with two chimeric proteins per transcript, and two of them are associated with three chimeric proteins per transcript. All chimeric proteins except two were modelled based on cases where an altProt-ORF overlaps a refProt-ORF. The only chimeric proteins that corresponded to overlaps between altProt-ORFs originate from transcript MtrunA17_Chr5g0430341, which is an mRNA molecule. Another transcript, MtrunA17_MTg0490971, was unique in this group because it corresponds to a ncRNA molecule. The remaining six transcripts were categorized as mRNA.

Chimeric proteins modelled with transcript MtrunA17_Chr1g0200071 can potentially prove the mosaic translation hypothesis if their products can subsequently be

detected by an antibody raised to a corresponding synthetic mosaic protein. In this candidate transcript, translation starts from the second (refProt frame), and a ribosomal frameshift changes the reading frame to the third (altProt frame). After the translation on the third frame, the ribosome changes the reading frame one more time from the third frame (altProt frame) to the second frame (refProt frame). It is also visualized in Figure 3.8. According to this figure, the first ribosomal frameshift occurs on the 860th base as a -2 frameshift, which corresponds to a backward movement of the ribosome by two nt. The second ribosomal frameshift occurs on the 942^{nt} base as a +2 frameshift, which is the slipping of the ribosome in the forward direction by two nt.

Chimeric proteins modelled with MtrunA17_Chr3g0144151, MtrunA17_ Chr4g0016371, and MtrunA17_MTg0490471 transcripts cannot be evidence for the mosaic translation. Both frameshifts found within these transcripts are of the same type, which indicates the two chimeric proteins cannot be part of a continuous protein sequence unless additional chimeric proteins are identified between them. For instance, in the MtrunA17_Chr4g0016371 transcript, the ribosome shifts twice from the third frame to the first frame, and both frameshifts are from an altProt-ORF to a refProt-ORF. Nevertheless, this transcript and the other transcripts exhibiting the same situation, can be examples of sequences that are associated with more than one ribosomal frameshifting.

Likewise, chimeric proteins modelled with transcript MtrunA17_Chr5g0430341 cannot be considered as direct evidence for mosaic translation unless further chimeric proteins are validated between them. Chimeric proteins on the transcript are visualized in Figure 3.9. Translation starts from the third frame, and a -2 ribosomal frameshift at the 1,134th base changes the reading to the first frame. The second ribosomal frameshift was validated at the 2,511th base as +2 frameshift from frame three to frame two. Because the first ribosomal frameshift brings translation to the first frame, and the second ribosomal frameshift starts from the third frame instead of the first one, an additional validated chimeric protein combining frames one and three must be found in support of the mosaic nature of this protein. Still, as mentioned above, the association of two ribosomal frameshifting events with one transcript is novel per se and should be followed in dedicated studies.

The second mosaic protein was validated by chimeric proteins modelled with MtrunA17_Chr6g0457461. Chimeric proteins on the transcript are visualized in Figure 3.10. Here, translation starts at the second frame, and a -2 ribosomal frameshift occurs at the 488th or 491st base, which changes the reading to frame three. Then, after translation on the third frame, another ribosomal frameshift occurs at the 552nd base as a +2 frameshift so that the reading frame changes from the third to the second frame.

Chimeric proteins modelled with MtrunA17_MTg0490971 could be evidence for mosaic translation if more ribosomal frameshifting positions were detected. The gene of this transcript is located on the mitochondrial chromosome. It is a member of the ncRNA group, and two frameshifts were detected on this transcript. Chimeric proteins on the transcript are visualized in Figure 3.11. In this example, translation starts from the first frame, and a +1 ribosomal frameshift changes the reading to the second frame at the 475th base. Then, after translation on the second frame, another ribosomal frameshift occurs at the 1,673rd base so that the reading frame changes from the second to the third frame. However, frame 2 between validated chimeric proteins modelled with MtrunA17_MTg0490971 has several stop codons. Extra evidence is needed to join them into a single continuous protein.

Validated mosaic proteins, namely associated with transcripts MtrunA17_Chr6g0457461 (and if validated later, also MtrunA17_MTg0490971), correspond to a scenario we call a “short round trip” in our earlier work (Çakır et al., 2021). This is a situation where the ribosome is brought back to the original reading frame by the second frameshift. Intriguingly, both cases illustrate the same type of double shift: frame two – frame three – frame two. In contrast, transcript MtrunA17_MTg0490971 (and if validated later, also MtrunA17_Chr5g0430341) exemplifies a situation we call a “one-way trip”, where the ribosome does not come back to the original frame after the last frameshift. Note that Figure 3.8 to Figure 3.11 were generated using Geneious (v. 7.1) software created by Biomatters, available from http://www.geneious.com.

Table 3.12. Transcripts that are associated with more than one chimeric protein modelled with conserved altProts.

Transcript IDs	Chimeric Protein IDs
MtrunA17_Chr1g0200071	(1) MtrunA17_Chr1g0200071_3F_828-1001_174_MtrunA17_Chr1g0200071_2F_218-1567_1350_-2_iteration_11_Within_5'_of_altprot_1540_Chimeric (2) MtrunA17_Chr1g0200071_3F_828-1001_174_MtrunA17_Chr1g0200071_2F_218-1567_1350_+2_iteration_1_Within_3'_of_altprot_1593_Chimeric
MtrunA17_Chr3g0144151	(1) MtrunA17_Chr3g0144151_1F_1222-1311_90_MtrunA17_Chr3g0144151_3F_33-1415_1383_-1_iteration_5_Within_3'_of_altprot_8319_Chimeric (2) MtrunA17_Chr3g0144151_1F_1222-1311_90_MtrunA17_Chr3g0144151_3F_33-1415_1383_-1_iteration_6_Within_3'_of_altprot_8320_Chimeric (3) MtrunA17_Chr3g0144151_1F_1222-1311_90_MtrunA17_Chr3g0144151_3F_33-1415_1383_+2_iteration_3_Within_3'_of_altprot_8338_Chimeric
MtrunA17_Chr4g0016371	(1) MtrunA17_Chr4g0016371_3F_798-1025_228_MtrunA17_Chr4g0016371_1F_1-1353_1353_-2_iteration_13_Within_3'_of_altprot_118439_Chimeric (2) MtrunA17_Chr4g0016371_3F_798-1025_228_MtrunA17_Chr4g0016371_1F_1-1353_1353_-2_iteration_20_Within_3'_of_altprot_118446_Chimeric
MtrunA17_Chr5g0430341	(1) MtrunA17_Chr5g0430341_3F_1059-1172_114_MtrunA17_Chr5g0430341_1F_1-1620_1620_-2_iteration_8_Within_3'_of_altprot_188974_Chimeric (2) MtrunA17_Chr5g0430341_2F_2501-2623_123_MtrunA17_Chr5g0430341_3F_2298-2531_234_+2_iteration_3_3'UTR_overlapped_with_CDS_188915_Chimeric
MtrunA17_Chr6g0457461	(1) MtrunA17_Chr6g0457461_3F_438-596_159_MtrunA17_Chr6g0457461_2F_170-715_546_-2_iteration_17_Within_5'_of_altprot_6844_Chimeric (2) MtrunA17_Chr6g0457461_3F_438-596_159_MtrunA17_Chr6g0457461_2F_170-715_546_-2_iteration_18_Within_5'_of_altprot_6845_Chimeric (3) MtrunA17_Chr6g0457461_3F_438-596_159_MtrunA17_Chr6g0457461_2F_170-715_546_+2_iteration_6_Within_3'_of_altprot_6896_Chimeric
MtrunA17_MTg0490471	(1) MtrunA17_MTg0490471_2F_422-538_117_MtrunA17_MTg0490471_1F_220-1740_1521_+1_iteration_16_Within_5'_of_altprot_299661_Chimeric (2) MtrunA17_MTg0490471_2F_1031-1147_117_MtrunA17_MTg0490471_1F_220-1740_1521_+1_iteration_7_Within_5'_of_altprot_299568_Chimeric
MtrunA17_MTg0490971	(1) MtrunA17_MTg0490971_2F_1562-1699_138_MtrunA17_MTg0490971_3F_1542-1727_186_+1_iteration_12_Within_3'_of_altprot_311571_Chimeric (2) MtrunA17_MTg0490971_2F_476-628_153_MtrunA17_MTg0490971_1F_277-534_258_+1_iteration_0_3'UTR_overlapped_with_CDS_312023_Chimeric

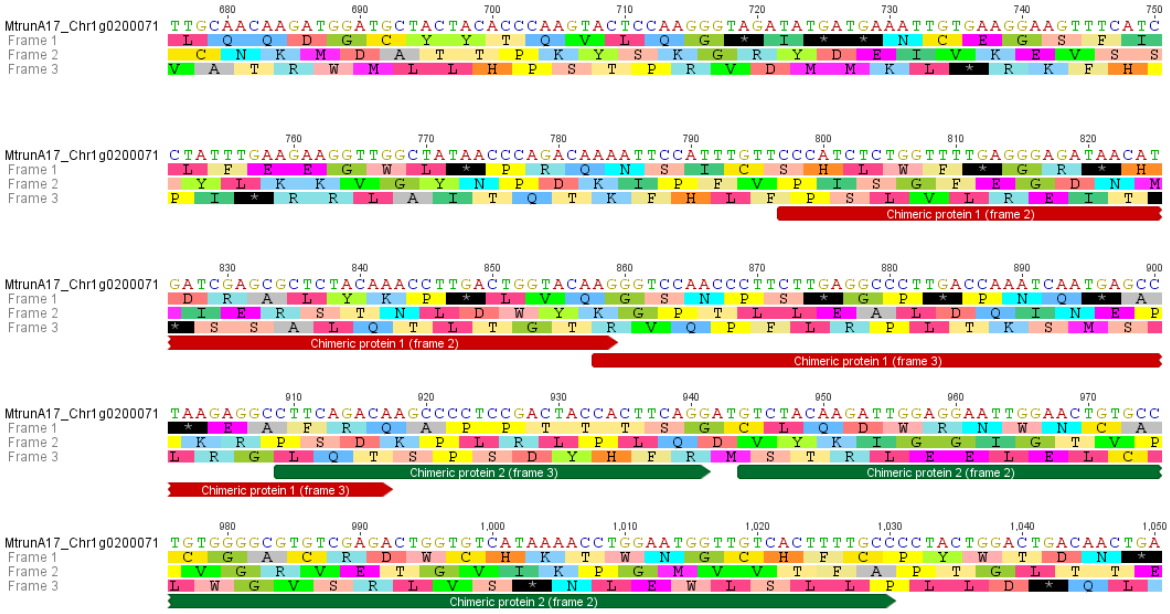

Figure 3.8. Mosaic translation on MtrunA17_Chr1g0200071. Two ribosomal frameshift positions were detected: 860th base as a -2 frameshift and 942nt base as a +2 frameshift.

Figure 3.9. Chimeric proteins on MtrunA17_Chr5g0430341 could be evidence for mosaic translation only if further chimeric proteins are validated between them. Two ribosomal frameshift positions were detected: around 1130th base as a -2 frameshift and 2510nt base as a +2 frameshift.

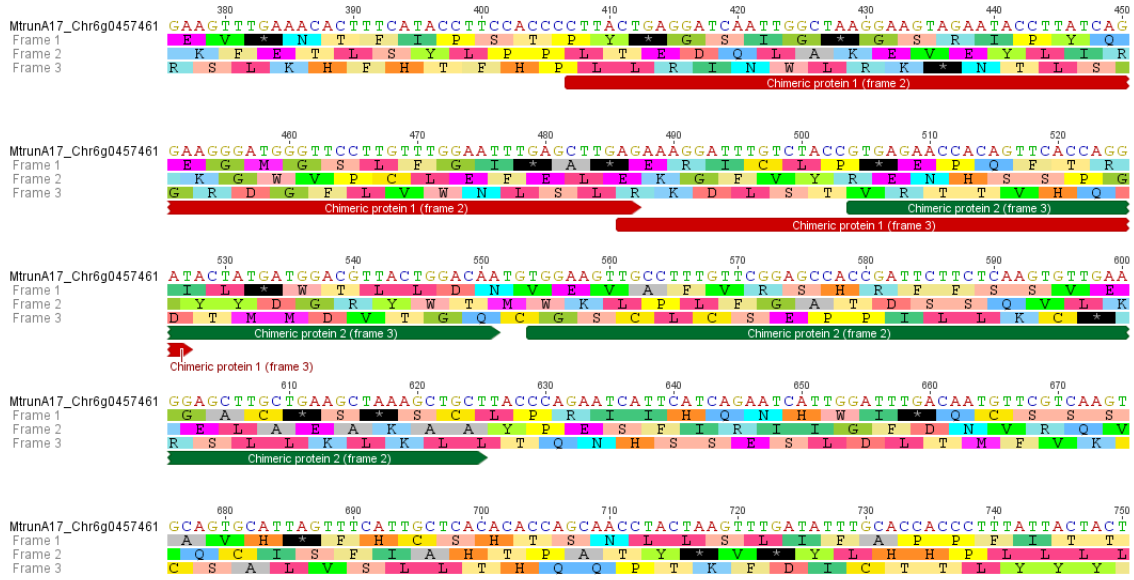

Figure 3.10. Mosaic translation on MtrunA17_Chr6g0457461. Three ribosomal frameshift positions were detected: 488th base and 491st as a -2 frameshift (frameshift on 491st base is not shown) and 552nd base as a +2 frameshift.

Figure 3.11. Chimeric proteins on MtrunA17_MTg0490971 could be evidence for mosaic translation only if further chimeric proteins are validated between them. Two ribosomal frameshift positions were detected: around 470th base as a +1 frameshift and 1670^{nt} base as a +1 frameshift.

3.6 Conserved AltProts and MS-Supported AltProts Found in *M. Truncatula* Genes Characterized So Far

How many altProts detected in our study are associated with genes functionally characterized in *M. truncatula*? To answer this question, we conducted a nearly comprehensive literature search for mRNA-type genes studied using at least one loss-of-function method. We also extended this search to ncRNA genes functionally analysed using any direct or indirect method. This article collection covered the years 1995-2022 and listed 325 genes: 293 mRNA-genes and 32 ncRNA-genes. Among them, five genes contained altORFs of MS-supported altProts identified in our study (four mRNA-genes and one ncRNA gene, one altProt per gene). The remaining 54 genes contained altORFs corresponding to conserved altProts identified in our study (Table 3.13, Table G.1 in APPENDIX G, and these tables with references, see section 2.7 Data Availability). In this context, the term "conserved" refers to the significant similarity (% identity) of an altProt to any annotated protein from the UniProt database (e-value equal to or below 0.001, the 70% identity threshold not applied). Surprisingly, this list contained many prominent genes involved in the root nodule symbiosis (48 genes) and equally well-known regulators of other biological processes (11 genes).

Among genes with MS-supported altProts, we found the GRAS-family transcription factor SCARECROW (SCR), which is essential not only for the rhizobial infection and root nodule number but also for root radial patterning and shoot gravitropism (Dong et al., 2021). Three other genes from this group were also essential for symbiosis with rhizobia: (1) nodule cysteine-rich protein 169 (NCR169), which is an ncRNA-encoded short peptide (Domonkos et al., 2013; Farkas et al., 2014; Horvath et al., 2015; Starker et al., 2006); (2) homo-glutathione (hGSH) synthase b (hGSHSb) involved in the control of nodule number (Frendo et al., 2001, 2005); and (3) cold acclimation specific 31 (CAS31), which is a dehydrin with a conditional nodulation phenotype (X. Li et al., 2018). The only non-symbiotic gene in this group was the HD-ZIPIII-family transcription factor revoluta 1 (REV1), which is essential for the leaf adaxial identity (C. Zhou et al., 2019). Among genes that contained altORFs with conserved altProts, we have found eight mRNA-genes essential for the fungal arbuscular mycorrhiza symbiosis. The majority of other genes in this group were SNF-related. They included two prominent nodule-related

membrane transporters characterized by our consortium: (1) multidrug and toxic compound extrusion 67 (MATE67), an iron-activated citric acid exporter essential for iron homeostasis in nodules (Kryvoruchko et al., 2018) and (2) natural resistance-associated macrophage protein 1 (NRamp1), an iron transporter essential for the nutrition of rhizobia-infected nodule cells (Tejada-Jiménez et al., 2015). Many other prominent SNF-related genes characterized by other groups were found in this category; for example, cystathionine beta-synthase 1 (CBS1), essential for the control of nodule number, had two corresponding conserved altProts with top hits of 75% and 98% identity (Sinharoy et al., 2013). A few SNF-related genes contained altORFs for more than two conserved altProts. Nodule cysteine-rich protein 247 (NCR247) had three corresponding conserved altProts with top hits between 61% and 73% identity (Farkas et al., 2014; Van De Velde et al., 2010). Intriguingly, they belonged to three different reading frames. The most interesting gene in this respect was an E2 ligase phosphate2-like (PHO2-like) essential for the control of nodule number (Curtin et al., 2017). Its transcript contained ORFs for five conserved altProts with top hit % identity ranging between 84 and 100. Again, they were found in all the three reading frames of this gene. The number of hits associated with these conserved altProts was between 41 and 14113, which is in contrast to most other altProts having only one or two hits each.

Whereas all altProts associated with published mRNA-genes are novel because their translation has not been studied by other groups, ncRNA-genes listed in Table 3.13 have been specifically targeted for translation of short ORFs. For that reason, it was important to understand if conserved and/or translated altProts identified in our study corresponded to the characterized short ORFs of those transcripts. We have found two such altProts: one MS-supported altProts corresponded to MtNCR169 mentioned above (Domonkos et al., 2013; Farkas et al., 2014; Horvath et al., 2015; Starker et al., 2006) and one conserved altProt corresponded to another member of NCR family MtNCR211, which is also required for the root nodule symbiosis (M. Kim et al., 2015; Starker et al., 2006). All other ncRNA-derived altProts identified in this study are novel. To the best of our knowledge, the possibility of their translation or biological reasons for their conservation have never been studied. It should be noted that altProts that have top hits with 100% identity may theoretically correspond to products of overlapping protein-coding genes. To rule out such a possibility, we manually checked each of the nine such cases listed in Table

G.1 in the *M. truncatula* genome browser (Pecrix et al., 2018). This brief analysis revealed that apart from MtNCR169 and MtNCR211, only one locus mentioned in this table had an overlapping protein-coding gene. Namely, MtENOD40-1 (C. Charon et al., 1997; Wan et al., 2007), which is annotated as a ncRNA-gene and overlaps with a hypothetical mRNA-gene MtrunA17_Chr8g0368434. This indicates that the remaining six altProts with 100%-identity top hits correspond to some evolutionary young (MtKNOX5, MtCLE34, and MtMATE1, one hit each) and some ultra-conserved (MtPHO2-like, MtLYK3, and MtDefMd1, between 151 and 14113 hits each) segments possibly translocated from other genomic locations. For references on these genes and details of the % identity values, see Table G.1. Such a high degree of conservation must be of great interest for deeper studies on these genes. Likewise, evolutionarily weakly conserved segments transferred from other genetic loci may also participate in the evolution of gene's function. Thus, ORFs of such altProts deserve further research regardless of the MS-based evidence for their translation.

Table 3.13. Characterized *M. truncatula* genes for which conserved altProts and MS-supported altProts were found in this study.

	Gene Symbol	**Medicago Gene ID v5**	**Transcript Type**	**Biological Process**	**Number of AltProts**
1	MtPIN3	MtrunA17_Chr1g0160461	mRNA	SNF	1 conserved
2	MtSYT3	MtrunA17_Chr1g0199571	mRNA	SNF	1 conserved
3	MtARF3	MtrunA17_Chr2g0282961	mRNA	SNF	1 conserved
4	MtCYP72A67	MtrunA17_Chr2g0288661	mRNA	SNF; saponin metabolism	1 conserved
5	MtVAMP721d	MtrunA17_Chr2g0291651	mRNA	SNF; AM symbiosis	1 conserved
6	MtCNGC15c	MtrunA17_Chr2g0326871	mRNA	SNF; AM symbiosis	1 conserved
7	MtPLT1	MtrunA17_Chr2g0328971	mRNA	SNF	1 conserved
8	MtYSL7	MtrunA17_Chr3g0109311	mRNA	SNF	1 conserved
9	MtGS1b	MtrunA17_Chr3g0110261	mRNA	SNF	1 conserved
10	MtNRAMP1	MtrunA17_Chr3g0124971	mRNA	SNF	2 conserved
11	MtYSL3	MtrunA17_Chr3g0127441	mRNA	SNF	1 conserved
12	MtKNOX5	MtrunA17_Chr3g0137241	mRNA	SNF	1 conserved
13	MtABCG59	MtrunA17_Chr3g0138261	mRNA	SNF; AM symbiosis	2 conserved
14	MtNLP1	MtrunA17_Chr3g0143921	mRNA	SNF	1 conserved
15	MtGbeta1	MtrunA17_Chr3g0144511	mRNA	SNF	2 conserved
16	MtFPN2	MtrunA17_Chr4g0004871	mRNA	SNF	1 conserved

Table 3.13. Characterized *M. truncatula* genes for which conserved altProts and MS-supported altProts were found in this study. (cont.)

	Gene Symbol	Medicago Gene ID v5	Transcript Type	Biological Process	Number of AltProts
17	MtP5CS3	MtrunA17_Chr4g0008951	mRNA	SNF; salt stress; drought	1 conserved
18	MtPHO2-like	MtrunA17_Chr4g0009054	mRNA	SNF	5 conserved
19	MtCNGC15b	MtrunA17_Chr4g0028861	mRNA	SNF; AM symbiosis	1 conserved
20	MtRab7a2	MtrunA17_Chr4g0034871	mRNA	SNF	1 conserved
21	MtVAMP721e	MtrunA17_Chr4g0043521	mRNA	SNF; AM symbiosis	1 conserved
22	MtRIT	MtrunA17_Chr4g0043744	mRNA	SNF	1 conserved
23	MtAKT1	MtrunA17_Chr4g0063141	mRNA	SNF	1 conserved
24	MtSUCS1	MtrunA17_Chr4g0070011	mRNA	SNF	1 MS-supported chimeric and 1 conserved
25	MtHAN1	MtrunA17_Chr5g0404131	mRNA	SNF	1 conserved
26	MtLYK3	MtrunA17_Chr5g0439631	mRNA	SNF	1 conserved
27	MtCBS1	MtrunA17_Chr6g0469911	mRNA	SNF	2 conserved
28	MtPIN4	MtrunA17_Chr6g0478431	mRNA	SNF	1 conserved
29	MtGS1a	MtrunA17_Chr6g0479141	mRNA	SNF	1 conserved
30	MtCAS31	MtrunA17_Chr6g0484671	mRNA	SNF	1 MS-supported
31	MtP5CS2	MtrunA17_Chr7g0239721	mRNA	SNF; salt stress; drought	1 conserved
32	MtLAX2	MtrunA17_Chr7g0241841	mRNA	SNF	1 conserved
33	MtSCR	MtrunA17_Chr7g0245601	mRNA	SNF; root development; shoot development	1 MS-supported
34	MtABCG56	MtrunA17_Chr7g0261971	mRNA	SNF	1 conserved
35	MtMCA8	MtrunA17_Chr7g0263361	mRNA	SNF; AM symbiosis	1 conserved
36	MthGSHSb	MtrunA17_Chr7g0273141	mRNA	SNF	1 MS-supported conserved and 1 conserved
37	MtNSP1	MtrunA17_Chr8g0344101	mRNA	SNF; AM symbiosis	1 conserved
38	MtMATE67	MtrunA17_Chr8g0352151	mRNA	SNF	1 conserved
39	MtARP3	MtrunA17_Chr8g0381261	mRNA	SNF	1 conserved
40	MtSymREM1	MtrunA17_Chr8g0386521	mRNA	SNF	1 conserved
41	MtNCR055	MtrunA17_Chr1g0166851	ncRNA	SNF	2 conserved;

Table 3.13. Characterized *M. truncatula* genes for which conserved altProts and MS-supported altProts were found in this study. (cont.)

	Gene Symbol	Medicago Gene ID v5	Transcript Type	Biological Process	Number of AltProts
					not MtNCR055
42	MtCLE34	MtrunA17_Chr2g0325371	ncRNA	SNF	1 conserved; not MtCLE34
43	MtNCR035	MtrunA17_Chr4g0007841	ncRNA	SNF	1 conserved; not MtNCR035
44	MtNCR211	MtrunA17_Chr4g0018031	ncRNA	SNF	2 conserved; one of them MtNCR211
45	MtNCR247	MtrunA17_Chr5g0423671	ncRNA	SNF	3 conserved; not MtNCR247
46	MtNCR044	MtrunA17_Chr7g0216231	ncRNA	SNF	1 conserved; not MtNCR044
47	MtNCR169	MtrunA17_Chr7g0229931	ncRNA	SNF	1 MS-supported conserved; MtNCR169
48	MtENOD40-1	MtrunA17_Chr8g0368441	ncRNA	SNF	1 conserved; not MtENOD40-1
49	MtSERF1	MtrunA17_Chr1g0170471	mRNA	Embryogenesis	1 conserved
50	MtAGa	MtrunA17_Chr2g0284911	mRNA	Flower development	1 conserved
51	MtREV1	MtrunA17_Chr2g0326731	mRNA	Leaf development	1 MS-supported and 1 conserved
52	MtMATE66	MtrunA17_Chr2g0328761	mRNA	Al3+ tolerance; Fe homeostasis	2 conserved
53	MtCCR1	MtrunA17_Chr2g0333781	mRNA	Lignin metabolism; stem, leaf, and flower development	1 conserved
54	MtMATE1	MtrunA17_Chr5g0442331	mRNA	Flavonoid metabolism; seed composition	1 conserved
55	MtPIN10	MtrunA17_Chr7g0255941	mRNA	Leaf development; cotyledon development; flower development	1 conserved

Table 3.13. Characterized *M. truncatula* genes for which conserved altProts and MS-supported altProts were found in this study. (cont.)

	Gene Symbol	**Medicago Gene ID v5**	**Transcript Type**	**Biological Process**	**Number of AltProts**
56	MtDefMd1	MtrunA17_Chr8g0339711	mRNA	AM symbiosis	1 conserved
57	MtAGb	MtrunA17_Chr8g0380021	mRNA	Flower development	1 conserved
58	MtLHA	MtrunA17_Chr8g0388921	mRNA	Saponin metabolism	1 conserved
59	MtSTF	MtrunA17_Chr8g0392991	mRNA	Leaf development	1 conserved

4. DISCUSSION

DNA corresponding to transcripts has coding potential and can encode functional proteins in all reading frames. Genome sequencing projects typically annotate the longest ORF in each transcript, and it is called a refORF or CDS, while other ORFs in the same transcript may or may not have the coding potential. If ORFs other than refORFs in a single transcript are predicted to encode proteins, they are referred to as altORFs. Proteins translated from altORFs are termed altProts. All theoretical altProts *in silico* translated using *M. truncatula* transcriptome data were analysed in various aspects, especially by MS and conservation analysis. We detected ~13,000 altProts that have similarities to the reference proteome database and 715 altProts with translation validated by MS analysis. Translation products of different reading frames can be combined in a single continuous polypeptide, which is possible by ribosomal frameshifting. These polypeptides are called chimeric proteins and mosaic proteins. To validate chimeric proteins and mosaic proteins, altProts can be used because they may represent building blocks for these proteins. In this project, chimeric proteins were modelled on the basis of overlapping ORFs, and their translation was validated by MS searches. As a result, 31 chimeric proteins modelled with MS-validated altProts and 116 chimeric proteins modelled with conserved altProts were validated. One chimeric protein was validated both in chimeric protein that was modelled by MS-validated and conserved altProts analysis, and, in total, 146 unique chimeric protein sequences were validated by MS. Finally, we found three mosaic proteins modelled with conserved altProts and produced by two ribosomal frameshift events each.

The proteome is the complete set of proteins expressed by an organism. This term can also be used to describe the assortment of proteins produced at a specific time in a particular cell or tissue type. The detection of all proteins expressed by the cell is technically challenging. Nevertheless, it is important to discover yet-unknown proteins and characterize hypothetical ones because knowledge of this hidden dimension of the proteome will transform our understanding of biological processes relevant to biomedicine, biotechnology, and agriculture. For instance, if currently unknown proteins have a function in the nodule formation and are upregulated during the nodule symbiosis, the complete understanding of the genetic basis of the nodulation process will be revolutionised. Over

the last 50 years, a considerable amount of information about the community of proteins has been uncovered. Still, the cell's whole proteome remains largely unknown. This portion of the proteome that has escaped identification for various biological and technical reasons is called the dark proteome (Laura Howes, 2022).

Proteome complexity and diversity are generated by processes that operate at the RNA and protein levels. At the RNA level, the major contribution to the proteome complexity and diversity is made by alternative splicing and mRNA editing. At the protein level, co- and post-translational modifications, alternative initiation sites, and peptide splicing by the proteasome further enhance proteome complexity and diversity. The formation of protein complexes is another contributing factor. Protein subunits can be joined into multiple configurations to create a series of protein assemblies with different functionalities (Chorev et al., 2015).

In this project, we elucidate one more mechanism operating at the translational level to further explain proteome complexity. After splicing, the mature mRNA is transported to the cytoplasm for translation. Ribosomes use the information carried by mRNA molecules to synthesize proteins. An ORF is a portion of nucleotide sequences that encodes a protein. It is commonly used to find protein-coding genes. Apart from the refORF, other ORFs, called altORFs, may encode proteins but are overlooked in most genome annotation projects (Raj et al., 2016; H. Xu et al., 2010). Proteins can be translated from altORFs in addition to canonical refORFs. Thus, a particular region on the same transcript can code up to three different proteins since three reading frames are present on the transcript. RNA is single-stranded, but DNA is double-stranded; thus, a particular DNA region can encode up to six different proteins. Consequently, a point mutation has the ability to change the amino acid composition of up to six proteins, a fact broadly ignored in large-scale genetic studies where synonymous mutations (synonymous with regard to the main or reference protein) are seldom considered as potential causes of mutant phenotypes.

In addition to translation from altORFs, a single polypeptide can incorporate products of more than one reading frame by ribosomal frameshifting. If one frameshift occurs during translation, the translated polypeptide is called chimeric protein because it

combines amino acid sequences of two reading frames. Furthermore, more than one frameshift during translation is possible. In such case, the translated polypeptide is called mosaic protein, and this mode of translation is called mosaic translation (Çakır et al., 2021). A mosaic protein may be composed of up to three reading frames. For instance, the translation starts at the first frame, and a ribosomal frameshift changes the frame from the first to the second. Then, another ribosomal frameshift on the same transcript changes the frame from the second to the third. This scenario is called one-way trip in our earlier report (Çakır et al., 2021). On the other hand, a mosaic protein may combine products of just two reading frames. For example, the translation starts at the third frame, and a ribosomal frameshift changes the frame from the third to the first. Then, another ribosomal frameshift brings translation back to the third frame, and this scenario is called the short round trip in our earlier report (Çakır et al., 2021). In this case, a mosaic protein is composed of the first and the third reading frame. In the present study, we based the identification of chimeric proteins on automatically generated model sequences corresponding to a wide range of possible ribosomal frameshifts. Each model sequence was used as a query in the search for exactly matching MS peptides. For technical reasons (inflation of the MS peptide search database), we limited the modelling of ribosomal frameshifting events to only the four most common types, namely +1, +2, -1, and -2, which correspond to movements by one or two nucleotides either in the forward or in the backward direction, respectively. An exception was made for altProt-ORFs located in UTR regions. For chimeric proteins modelled with UTR-located altProt-ORFs, frameshifting events of the longer distance (up to 10 nucleotides in either direction) were considered. Longer ribosomal frameshifts such as four, five, and six nucleotide shifts were reported in the literature (Caliskan et al., 2017; Weiss et al., 1987; Yan et al., 2015).

RNA carries out an extensive range of functions; for instance, while mRNAs are coding RNAs translated into protein by the ribosome, other RNAs are thought to have no coding functions: tRNAs transport amino acids to ribosomes as they synthesize proteins, rRNAs combine with proteins to form the ribosomes, microRNAs affect gene expression especially important in growth and development. Many new types of non-coding RNAs have been discovered recently, and their roles have been verified in diverse biological processes (Sun & Chen, 2020; Vazquez-Anderson & Contreras, 2013). However, in our study, altORFs and their translation products altProts were not limited to mRNA

transcripts; ncRNA, rRNA, and tRNA transcripts were analysed to determine ORFs translated into proteins. The latter two groups were tested for translation as other studies found at least six functional proteins translated from rRNA (Root-Bernstein & Root-Bernstein, 2016), while tRNA was found to be associated with many conserved ORFs in our preliminary analysis. The ncRNAs were included as some non-canonical protein-coding transcripts can be wrongly classified as ncRNAs. For example, many pseudogenes and transcripts coding for short peptides were initially placed into this category (Cheetham et al., 2020; Zlotorynski, 2020). Eukaryotic organisms such as plants and animals contain thousands of ncRNAs. They are generally thought to lack ORFs and protein-coding potential. However, through the development of Ribo-Seq and other sequencing technologies, an increasing number of studies, especially cancer studies, have shown that ncRNAs are translated (B. Zhou et al., 2021).

ORFs in non-mRNA transcripts are typically ignored because the minimum ORF length threshold is too high in genome annotation projects (up to 300 nucleotides). Another reason for the exclusion of unusual ORFs is related to the ORF definition. According to the classical definition, ORFs must have a start codon AUG and be at least 300 nt in length, and this definition is adopted in most genome annotation projects (Benitez-Cantos et al., 2020; Steward et al., 2017; Yazhini, 2018) To reveal the unknown protein-coding potential of each transcript group, we checked all ORFs using a lower length threshold of 60 nt. We also included ORFs starting with any sense codon regardless of the presence of AUG. Whereas rRNA and tRNA transcripts were included in our analysis as we considered they may have a dual function and be translated into functional proteins, miRNA was excluded from analysis because all ORFs in miRNA transcripts are shorter than the minimum length threshold of 60 nt. Even though we validated many translated ncRNA-derived altProts, we found no evidence for translation either from rRNA or tRNA-derived altORFs. This endeavour could be successful if we used the human proteome, which has a much larger database of MS-derived peptides. Extending this type of analysis to the human proteome may be very informative and can deliver many unexpected discoveries. So far, rRNA and tRNA are not included in the largest database of alternative proteins in humans (Brunet et al., 2021)

AltProts may act as building blocks for chimeric proteins and mosaic proteins; thus, altProts can be used to demonstrate the existence of these frameshifted proteins. That is, chimeric or mosaic proteins are composed of different altProts on the same transcript; thus, validation of altProts is necessary for the validation of chimeric proteins and mosaic proteins. AltProt validation was conducted by two approaches: MS and conservation. In the first approach, all theoretical altProt sequences were validated by MS searches using SearchGUI and Peptide Shaker software. In the second approach, all theoretical altProts were compared to the reference protein database. If altProts are expressed, they may be conserved among other species (L. J. Jensen et al., 2003; Lockwood et al., 2019). At the same time, sequences found twice or more in the source species (*M. truncatula*) but absent from other organisms may reflect recent DNA translocation events, which are important in the evolution of genes. For this reason, we combined all altProts with external (different species) and internal (source species) conservation signatures in one group called here conserved altProts. DIAMOND, which is a high-throughput protein alignment tool, was used for conservation analysis as it is faster than the other commonly used alignment tool BLAST, and DIAMOND is optimized to handle a large number of queries (e.g. ~800,000 altProts) (Buchfink et al., 2014; Camacho et al., 2009). After the validation of altProts by MS or determination of conserved altProts, validated altProts that overlap with their refProts and other altProts in the same transcript were used for chimeric protein modelling. Such overlapping ORFs were subsequently used for the validation of mosaic proteins. In other words, if two ORFs overlap at the transcript level, these two ORFs may be translated into a single polypeptide joined by a ribosomal frameshift. Using our in-house script, all chimeric proteins corresponding to +1, +2, -1, and -2 frameshifts were modelled for each pair of overlapping ORFs to determine the exact ribosomal frameshift position. The length of modelled chimeric proteins was limited to 40 aa because the second frameshift may cause invalidation of the modelled chimeric proteins if it is very close to the first ribosomal frameshift position. Furthermore, the chimeric protein length limit of 40 aa was meant to facilitate the identification of mosaic proteins because we hypothesized that more than one ribosomal frameshift may occur during translation of a single transcript. Modelled chimeric proteins were searched and validated by MS. Afterwards mosaic proteins were validated independently if a transcript was found to be associated with more than one chimeric protein. This protocol was based on the notion that every chimeric protein corresponds to one ribosomal frameshifting event. Validated novel protein sequences,

altProts, chimeric proteins and mosaic proteins, can be submitted to the protein synthesis pipeline for further validation by antibodies, structural and functional analysis (see section 4.9 below). Please note that there is a possibility that although one transcript is associated with more than one chimeric protein, these chimeric proteins may be produced independently without being combined in a continuous polypeptide sequence. Wet-lab experiments should address this uncertainty using antibodies raised to synthetic mosaic proteins. Regardless of the unequivocal proof of mosaic nature of these proteins, the presence of multiple ribosomal frameshifting associated with a single transcript is so far unknown in non-viral genomes except the viral Gag-Pro-Pol polypeptide (Hatfield et al., 1992; Jacks, 1990), which makes these results very novel. Likewise, the identification of transcripts associated with single ribosomal frameshifting events is a significant discovery of our study since chimeric proteins were thought to be extremely rare in eukaryotes (Farabaugh, 2006; Ketteler, 2012). Finally, multiple translated and conserved altProts found by our approach constitute a unique resource for functional studies, which should be aimed at independent characterization of overlapping proteins and elucidating their potential interactions (Aspden et al., 2014). This resource is not limited to symbiosis-related genes and will be of high interest to a broad range of plant biologists.

4.1 Although Proteomics Technology is Developing Rapidly, Experimental Validation of Mosaic Translation is Still a Challenge

Many experimental methods were developed to identify novel proteins, the most common method to identify protein-coding regions in the genome is examining ORFs (Anders et al., 2021). Putative ORFs are usually 1,000 to 2,000 nt long, but translated ORFs range from less than 100 nt to more than 2000 nt (K. T. Jensen et al., 2006; Mir et al., 2012). Genome and/or transcriptome sequencing is necessary to determine all ORFs, while translated ORFs can be validated by various computational and experimental methods such as conservation evidence, MS analysis, and RIBO-Seq (Olexiouk & Menschaert, 2016; Zhu et al., 2018).

Genome sequencing, also known as whole genome sequencing (WGS), refers to sequencing the entire, or nearly the entire, genome of an organism at a single time and contains information on an organism’s chromosomal and mitochondrial DNA and, for

plants, DNA contained in the chloroplasts. WGS reads contain both coding and non-coding sequences (Schon et al., 2021). Genetic information is transferred from DNA to RNA through transcription, and transcriptome sequencing refers to sequencing all RNAs, or almost all RNAs, including coding and non-coding RNAs in an individual or a population of cells by cDNA sequencing (B. Wang et al., 2019).

The second-generation sequencing technology revolutionized genomic and transcriptomic analysis by increasing throughput and lowering costs. However, it uses only short reads, less than 300 bases. Thus, it is also called short-read sequencing. Short cDNA fragments between 50 to 300 bases in length are sequenced, limiting the detection of alternative splicing and protein isoforms. Moreover, the second-generation sequencing technology relies on reverse transcription and polymerase chain reaction, which cause bias and are error-prone. Thus, the detection of novel exon boundaries becomes problematic. In other words, the second-generation sequencing technology is unable to accurately detect novel exon boundaries. However, the third-generation sequencing technology relies on cheaper, faster, and more sophisticated processes (Heather & Chain, 2016). One advantage of third-generation sequencing is that a single molecule with ~10-18 Kb length can be sequenced with PacBio technology, and even a much higher sequencing length of 100 Kb can be reached with Oxford Nanopore sequencing technology (Guo, 2018). Thus, sequences of spliced transcripts and exon boundaries, which are necessary for the detection of chimeric proteins and mosaic proteins, can be more confidently detected by third-generation sequencing.

Although we identified and validated 715 altProts, 31 chimeric proteins modelled with MS-validated altProts, and 116 chimeric proteins modelled with conserved altProts, only two mosaic proteins were validated. These two mosaic proteins were modelled with conserved altProts. Of course, these unique proteins may be false positives, so they should be further validated by wet-lab experiments. Additionally, among 715 altProts, only 142 altProts were validated in more than one organ/condition. The remaining 573 altProts were validated for only one organ/condition and were called organ/condition-specific. They may indeed be translated only in a specific organ or condition. Furthermore, we analysed a dataset that is called whole plant, and if one altProt is expressed in a specific organ, theoretically, it should also be identified in the whole plant dataset. Still, these datasets

were generated by different groups at different times and conditions. In addition, the ability to detect an organ-specific altProt strongly depends on the abundance of the protein. Since in the whole plant dataset, organ-specific altProts are strongly diluted, their absence in this group is not a strong indicator of their false positive status.

AltProt sequences are usually shorter than refProts; thus, some altProts may be too similar to refProts in such a way that MS search algorithms cannot differentiate altProt peptides from refProts, and altProts may be grouped with refProts even though altProts are expressed. Then, these altProts could not be validated. Similarly, modelled chimeric proteins are 40 aa in length and can be too similar to canonical proteins. A significant portion of chimeric proteins could not be validated just because of the high similarity to canonical proteins. For this reason, many altProts and translated chimeric proteins may not be validated, and mosaic proteins corresponding to those non-validated sequences could not be validated afterwards, even if these mosaic proteins were indeed translated and have vital functions in the cell. To overcome this problem, conservation evidence was also used for translation. Even though some altProts are really translated, sometimes their presence cannot be verified by MS analysis due to the expression in a specific organ/condition. For instance, if an altProt is expressed under drought or salinity conditions, we cannot verify its translation from the data produced from the "normal" condition. During the sample preparation for MS analysis, some altProts maybe wash out. In addition, MS algorithms, in general, cannot verify the low-abundance altProts (Filip et al., 2015; Lu et al., 2011). For these reasons, we also compared altProt sequences to databases of known proteins. Namely, we checked the similarity of altProts to annotated protein from any organism.

Transcripts can be searched for sequences highly similar to known ribosomal frameshifting sites to detect chimeric proteins and mosaic proteins. Even though this *in silico* analysis can potentially be useful, it does not provide the whole spectrum of frameshifting signals because there are no universal known ribosomal frameshifting sites, and they may be species-specific. Furthermore, the minimum length between two adjacent frameshifting sites is unknown, making it challenging to model chimeric proteins properly because we want to model chimeric proteins without inflating the search database (Brierley et al., 2010). The problems explained above present a major obstacle to a large-scale identification and unequivocal validation of altProts, chimeric proteins, and mosaic

proteins. In addition, even if a researcher observes these kinds of proteins in a specific study, the dominating paradigms that deny the existence of such proteins prevent focusing on them as research targets. In most cases, they are considered to be artifacts having nothing to do with biological functions. In view of this neglect and the absence of alternative methods capable of detecting long continuous protein sequences of low abundance (a proteomic analogue of long-read Nanopore sequencing), our approach is the only large-scale method of mosaic protein identification available to the scientific community. Therefore, despite some uncertainties associated with our analysis, this study is pioneering in the field, and its results fully deserve wet-lab validation.

4.2 Unique Features of Ribosomal Frameshifting Make its Products Different from Proteins Produced by Other Mechanisms

Ribosomal frameshifting events are triggered by signals. The sequence of the transcript can trigger a frameshift, or sequence-independent mechanisms such as RNA secondary structure may also trigger frameshifts. Ribosomes may stall on a frameshift sequence, also known as slippery sequence, and these sequences are usually rare codons for which few tRNAs are available. This stall triggers the ribosome to change the reading frame by a ribosomal frameshift. However, a slippery sequence is specific to the organism, and most organisms' slippery sequences have not been researched yet. Additionally, many slippery sequences are known for viruses but are not conserved among viruses (Brierley et al., 2010; Gurvich et al., 2005; Kawakami et al., 1993; Korniy et al., 2019). Furthermore, RNA secondary structures such as pseudoknot, stem-loop, or kissing loop structures act as a roadblock to pause translation and trigger ribosomal frameshifting (Caliskan et al., 2014; Kima et al., 2014; Korniy et al., 2019).

The main advantage of mosaic translation is that more than one polypeptide can be produced from a single spliced transcript, so cells can change the protein content without transcribing new RNAs. For instance, using such mode of translation, cells may quickly produce different proteins from the already available transcripts in stress conditions. Cells can respond to environmental changes more rapidly by mosaic translation. Like other mechanisms that increase protein-coding capacity, such as alternative splicing, ribosomal frameshifting also increases the genome's protein-coding capacity. Additionally, ribosomal

frameshifting can control gene expression by influencing mRNA stability and regulating the stoichiometric ratio between proteins (Advani & Dinman, 2016; Atkins et al., 2016). It is known that ribosomal frameshifting is required for several human pathogenic viruses because certain viral enzymes are produced by a frame other than the reference frame. In addition, ribosomal frameshifting regulates the ratio between viral structural proteins, which is necessary for virion assembly (Korniy et al., 2019).

Frameshifted proteins, chimeric proteins and mosaic proteins, contribute to the protein-coding capacity of the genome just as alternative splicing and peptide splicing by the proteasome does. Ribosomal frameshifting is similar to alternative splicing but operates on spliced transcripts. Alternative splicing is one of Nature's inventions to diversify proteomes. It works by selecting different combinations of splicing sites within a precursor mRNA to synthesize differently spliced mRNAs. These variable spliced mRNAs can encode different proteins and have different functions and activities, but they all result from a single gene (Schwarzenbach, 2013). Because alternative splicing joins different combinations of exons, the reading frame may be changed for a particular RNA region of the transcript. For instance, one ORF is considered as a refORF, and another ORF in the same region is thought to be an altORF in one spliced transcript. However, in a differentially spliced transcript of the same gene, these ORFs may be considered as the exact opposite of their previous status: what was thought to be a refORF can become an altORF and the other way around. Thus, there is a possibility that validated altProts correspond to the unknown or undetected spliced transcripts. For that reason, ideally, studies on altProts should be accompanied by resequencing of spliced transcripts with long-read methods in reliable quality to avoid such a possibility. At the same time, even if all altProts validated using MS are products of alternative splicing, this does not diminish their novelty and importance for the interpretation of mutant phenotypes. Their discovery goes ahead of long-read-based transcriptomic studies because these hypothetical alternative isoforms have never been detected using conventional methods. Like in the case of altProts, their existence emphasizes the need to take into account mutations synonymous for refProts because they may change the amino acid sequences of alternative isoforms. Like altProts, these isoforms may have functions completely different from those attributed to refProts because the amino acid sequences of such isoforms are quite different from the annotated protein sequences of their transcripts. Thus, their comprehensive

characterization requires knocking out each isoform without affecting their refProts. In this sense, alternative isoforms that use ORFs other than refORFs require the same methodology as altProts. It should be noted that synonymous mutations are consistently excluded from studies that are based on forward genetics. Such studies are aimed at the identification of genetic loci responsible for a mutant phenotype. For example, a failure to recognize the link between a synonymous mutation and a genetic disorder in humans results in a major delay in making an accurate diagnosis and in administering an adequate treatment. Thus, our approach should be extended to the human transcriptome.

In contrast to altProts validated by MS, chimeric proteins and mosaic proteins are conceptually less likely to be produced by alternative splicing. The only scenario in which such ambiguity exists is the presence of sequences conserved at the protein level and present more than once throughout the length of the genomic region of a corresponding transcript. This, however, is easy to rule out for each chimeric protein and mosaic protein validated by MS. Naturally, we conducted such analysis for all chimeric proteins and mosaic proteins reported in our study. None of them can be produced by alternative splicing because sequences found on either side of the frame fusion are unique for their genomic DNA. This information was obtained from the detailed sequence comparison between chimeric proteins and three-frame translation sequences of their genomic DNA using TBLASTN (results not shown). Trans-splicing is another mechanism that can potentially mimic chimeric proteins and mosaic proteins. This mechanism joins parts of mRNA that either belong to different genetic loci or to antisense strands of the same loci (Lasda & Blumenthal, 2011). Trans-splicing has been reported in the genus Medicago (Z. shui He et al., 2008). However, it would be a great coincidence indeed to find that a DNA segment corresponding to a hypothetical but MS-supported ribosomal frameshifting site also corresponds 100% with a sequence produced by a hypothetical (and so far not documented) trans-splicing event. In this unlikely scenario, the sequence would be of even more interest for downstream analysis. Once long-read transcriptomic data become available for *M. truncatula*, we plan to address such a possibility using dedicated software called Genion, which is an accurate gene fusion caller (Karaoglanoglu et al., 2022).

Peptide splicing by the proteasome is another mechanism similar to ribosomal frameshifting. However, peptide splicing processes peptides produced from a single

reading frame. For that reason, products of peptide splicing cannot be mistaken for altProts, chimeric proteins, or mosaic proteins. In addition, peptide splicing differs from ribosomal frameshifting because it acts after the translation, while ribosomal frameshifting takes place on transcripts during translation. Peptide splicing operates by cleavage of a ready amino acid sequence at specific positions and subsequent stitching of its parts through a transpeptidation reaction either in the sequential or rearranged order (Vigneron et al., 2017).

4.3 Ribosomal Frameshifting May Involve Specialized Ribosomes

Translation control is increasingly recognized as a significant factor in determining protein levels. Previously, ribosomes were thought as rigid cellular machines that mediate protein synthesis. Their role in this process was considered essential but invariant because translational regulation was thought to be mediated by other auxiliary factors, and ribosome recruitment was the endpoint of the regulation according to this earlier view (Guo, 2018). However, recent developments in the last decade have revealed that heterogeneous types of ribosomes can be present in different tissues, and more importantly, these ribosomes can preferentially translate different subsets of mRNAs. These heterogeneous types of ribosomes, also called specialized ribosomes, translate different transcripts in different conditions (Dinman, 2016; Ferretti & Karbstein, 2019; Gilbert, 2011; Xue & Barna, 2012). Thus, we hypothesized that specialised ribosomes may be required for translation of altProts, chimeric proteins, and mosaic proteins (Çakır et al., 2021).

For over 30 years, it has been known that prokaryotic ribosomes act as sensors (Bischoff et al., 2014; Cheng-Guang & Gualerzi, 2021; VanBogelen & Neidhardt, 1990). Experimental support for this possibility in eukaryotic ribosomes has been obtained recently. The ribosome can act as a metabolite multi-sensor; for instance, some metabolites, like polyamine or sucrose, regulate gene translation (Van Der Horst et al., 2020). Their conceptual counterparts, spliceosomes, are known to receive and track signals that regulate spliceosome activity during RNA maturation (Y. Cao & Ma, 2019). Similarly, it seems theoretically plausible that ribosomes may respond to internal and external stimuli by altering the reading frames. This ability can be the basis for producing altProts,

chimeric proteins, and mosaic proteins precisely tailored to specific environmental conditions, developmental stages, cell types, etc. However, this idea does not exclude a possibility that environmentally controlled ribosomal frameshifting can be useful under standard conditions.

4.4 Does Translation Initiation Involve AUG in AltProts, Chimeric Proteins, and Mosaic Proteins?

Although many proteins have been identified to use non-canonical start codons (other than AUG) for translation initiation (Kearse & Wilusz, 2017; Ma et al., 2014), quantifying the presence of AUG in altProts, especially in conserved and MS-validated altProts, is informative as most genome annotation projects use this conventional start codon for determining translated ORFs (Benitez-Cantos et al., 2020). Among altProts that have top hits with at least 70% identity, ~8,100 (62%) altProts have at least one in-frame start codon AUG, and ~5,000 (38%) altProts have no in-frame start codon. Looking at individual types of RNA reveals that mRNA, ncRNA, rRNA, and tRNA-derived altProts with an in-frame start codon AUG constitute ~5,800 (67%), ~2,000 (56%), 174 (45%), and 135 (32%) of their respective transcript groups as shown in Table 4.1.

Among altProts that have in-frame start codons AUG and top hits with at least 70% identity, the relative position of the first in-frame start codon falls into the first half of the altProt's length in 70% of cases, as shown in Figure 4.1. This indicates that translation of corresponding altProts may be initiated even with a canonical AUG and result in relatively long proteins (at least one-half of the theoretical altProt's length). The relative position of the in-frame start codon per RNA type is shown in Figure E.1.

Table 4.1. Distribution of an in-frame start codon AUG among altProts with top hits of at least 70% identity

	Present		**Absent**		**Total**	
	N	**%**	**N**	**%**	**N**	**%**
mRNA	5,818	66.70%	2,900	33.30%	8,718	100.00%
ncRNA	1,968	55.50%	1,581	44.50%	3,549	100.00%
rRNA	174	45.20%	211	54.80%	385	100.00%
tRNA	135	31.70%	291	68.30%	426	100.00%
Total	8,095		4,983		13,078	

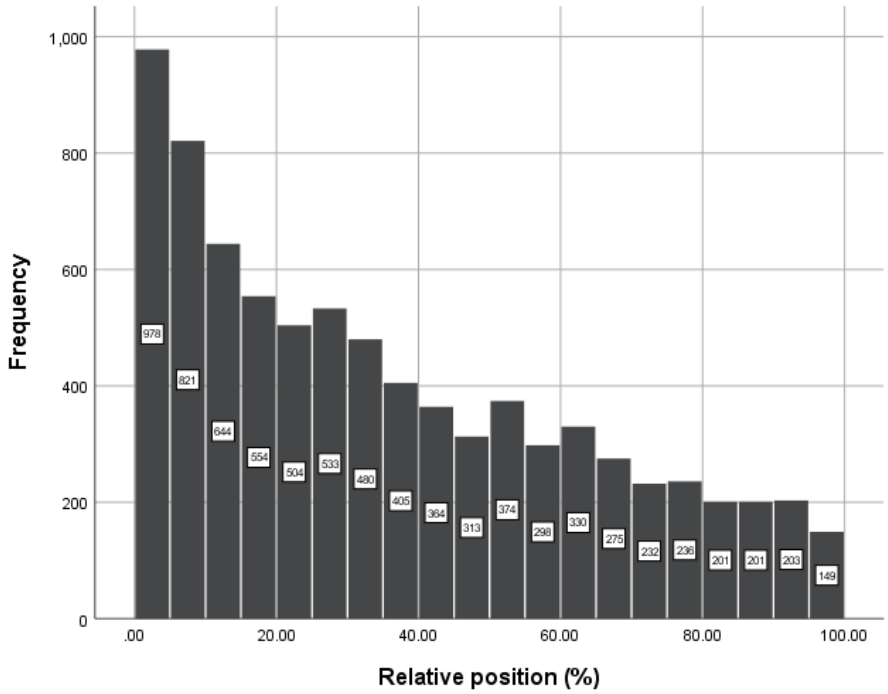

Figure 4.1. Relative position of an in-frame start codon AUG for altProts with at least 70% identity. $\bar{x}$ = is 35.8, SD = 27.6, n = 8,095. In 69.8% of altORFs, AUG is present in the first half of the altORF's sequence, which indicates a potential for the synthesis of relatively long proteins. The bin size was set to 5.

In the group of 715 altProts validated by MS searches, 538 (75%) of altProts have at least one in-frame start codon AUG, and 177 (25%) altProts have no in-frame start codon. Among the 538 altProts with at least one in-frame start codon, 485 altProts belong to the mRNA group, and the remaining 53 altProts belong to the ncRNA group. The first in-frame start codon is present in the first half of the altProts in 79% of those altProts, as shown in Figure 4.2. Thus, most of the conserved and MS-validated altProts identified in our study have an in-frame start codon AUG that is located in the first half of their length.

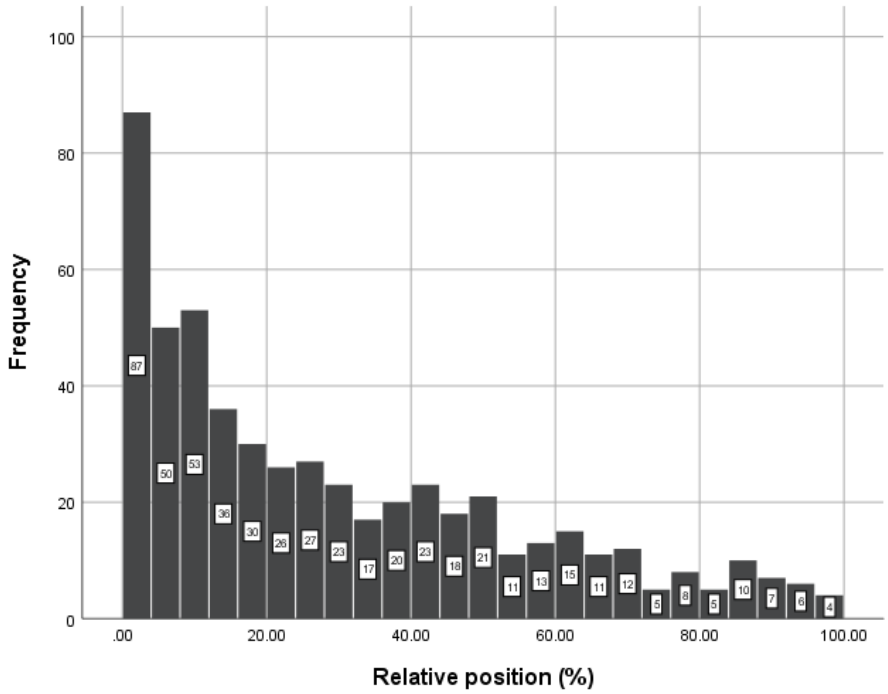

Figure 4.2. Relative position of an in-frame start codon AUG for MS-validated altProts. $\bar{x}$ = 29.3, SD = 25.9, n = 538. In 78.8% of altORFs, AUG is present in the first half of the altORF's sequence, which indicates a potential for the synthesis of relatively long proteins. The bin size was set to five.

4.5 Inclusion of All Possible Altprots Inflates the Search Database

Due to database inflation, the validation of all altProts by MS searches is usually impossible. Inclusion of all altProts increases the size of the search database resulting in limited and inefficient validation of altProts. Several methods are proposed to deal with inflated search databases, and each has advantages and disadvantages. These methods can be split into two categories: database dependent and database independent. In a database-dependent search, the amino acid sequence of proteins is determined with the assistance of a sequence database, so if a protein sequence is not present in the database, it is not validated. On the other hand, in a database-independent search, also called de novo peptide sequencing, the amino acid sequence of proteins is determined from the spectrum without a sequence database. The latter is considered to be a powerful method for large database searches. However, a database-dependent search is usually considered to be more efficient on the same dataset if the database size is not inflated (Johnson & Taylor, 2000; Muth & Renard, 2018; P. Wang & Wilson, 2013). Additionally, de novo sequencing is more error-

prone compared to a database-dependent search. Thus, we used a database-dependent approach with some modifications for a large database search, a procedure called two-step approach (Fu & Li, 2005; Muth & Renard, 2018).

All altProts and modelled chimeric proteins were searched in two publicly available datasets: PXD002692 (Marx et al., 2016) and PXD013606 (Shin et al., 2021). Translation capacity of all ORFs on all major types of RNA (mRNA, ncRNA, rRNA, and tRNA) was assessed. For the determination of all ORFs, a 20 aa minimum length threshold was chosen; that is, ORFs with less than 20 aa were not considered. mRNA-derived altProts were verified using a two-step MS approach. Non-mRNA-derived altProts (ncRNA, rRNA, and tRNA-derived) were analysed by a regular MS search, also called one-step procedure. The two-step MS approach consists of two steps: the first and the second step. In the first step, mRNA-derived altProts were split into 10 equal groups. Reference protein sequences, also known as refProts, and contaminant sequences were added to each group. Then, each group was searched independently. In the second step, validated altProts from each group were combined to generate a single search database, and refProts and contaminants were also added to this database, which was searched one more time in the same dataset. Note that searches in the PXD002692 and PXD013606 datasets were conducted independently, for instance, validated altProts from the first step of PXD002692 was not included in the second step of PXD013606, or vice versa. Besides, chimeric proteins modelled with conserved altProts were searched by the same approach, that is, two-step approach.

The reason why altProts identified from different datasets were not combined for the second step is that low-quality data may cause the validation of untranslated altProts. These are false-positive altProts, and the inclusion of these artefactual sequences may inflate the search database in the second step. Then, the number of validated true altProts may decrease, and false-positive altProts may be validated in the second step due to the low quality of one or several datasets. In other words, low-quality datasets may affect the validation of proteins, so the two-step approach can be used to avoid it. In this project, only two datasets were used. Thus, the importance of using the two-step approach may not be evident. However, altProts from other organisms, such as humans, may be searched for in hundreds of datasets. In such a case, low-quality datasets may affect the overall validation.

The two-step approach ensures that each dataset can be used to validate altProts regardless of previous knowledge, and low-quality datasets do not affect the validation of another dataset's results.

4.6 SearchGUI and PeptideShaker Are Computational Proteomics Tools for Validation and Quality Control

MS searches were conducted by SearchGUI (Barsnes & Vaudel, 2018) and its partner tool PeptideShaker (Vaudel et al., 2015), which are written in Java. SearchGUI has a user-friendly graphical interface for configuring and running proteomic search and *de novo* engines dedicated to database-independent searches. It currently supports the following engines: X!Tandem, MS-GF+, OMSSA, Comet, Andromeda, MyriMatch, MS Amanda, Tide, DirecTag, MetaMorpheus, and Novor. The inclusion of all search engines in proteomic data analysis may increase the search time drastically, and analysis of large proteomic datasets needs considerable computer resources. Thus, if possible, SearchGUI should be used by the command line for large proteomic data analyses. The command line tool for SearchGUI is SearchCLI, but the scientific community often refers to it as SearchGUI regardless of using it by command line. It reads MS^2 files in MGF or mzML, and database files should be in Fasta format (Kopczynski et al., 2017).

We preferred to use MGF file format that was converted from raw files via ThermoRawFileParser (v. 1.1.2) (Hulstaert et al., 2020). We did not notice any difference in the results using either MGF or mzML format (Deutsch, 2012). MGF file format is a text-based file format for mass spectrometer output files. It stores MS^2 spectra along with related meta information on the level of MS^2. The minimum definition of an MGF file is charge, precursor mass, and m/z - abundance pairs. Like MGF file, mzML file is XML-based and presents another commonly used MS output file format (L. He et al., 2015). Besides, SearchGUI with a graphical interface supports raw files and converts MGF or mzML file formats by msconvert (Adusumilli & Mallick, 2017). However, this option was only available in the graphical user interface due to license issues. The data conversion tool msconvert (Adusumilli & Mallick, 2017) or ThermoRawFileParser (Hulstaert et al., 2020) should be run separately from the command line for converting file formats. SearchGUI

creates all results in a single compressed file, a ZIP file, which can be forwarded to PeptideShaker (Kopczynski et al., 2017).

PeptideShaker is used for the interpretation of results that SearchGUI generates. PeptideShaker can combine identification results that are generated by different search engines. It can recalculate PTM localization scores and redo protein inference based on multiple search engines. Like SearchGUI, PeptideShaker can be used by graphical user interface as well as the command line. Like with SearchGUI, for large proteomic data analyses, PeptideShaker should be used with the command line to decrease running time. SearchGUI results in a ZIP format can be read by PeptideShaker automatically without any additional parameters such as the path location of MGF files, because the ZIP file contains all necessary information for data processing (Kopczynski et al., 2017).

On the other hand, additional adjustments to the parameters, such as changing FDR or FNR, are still possible in PeptideShaker. Different validation thresholds can be used on the same SearchGUI result. Furthermore, the graphical user interface provides various plots and tools, such as 3D structures of validated proteins (if a protein model is available, unknown proteins are not supported), quality control plots, gene ontology mapping and other useful features. Moreover, PeptideShaker uses the target-decoy approach for confidence results; that is, half of the search database contains proteins that we want to find, and the other half contains decoy sequences which are incorrect sequences and are not expected to be validated. Decoy sequences are the reverse aa sequences of proteins in the database. For instance, altProt, refProt, and contaminant databases were used as a search database in the altProt validation procedure, and their aa sequences were reverted to generate a decoy database. The target-decoy approach is used to estimate how many false positives are associated with the validated proteins, calculates a threshold to estimate FDR, and then filters those validations using the threshold (Farag et al., 2021; Z. Zhang et al., 2018). Then, validated proteins can be exported at the PSM, peptide, and protein levels, and additionally, they are exported in the user's custom report format (Kopczynski et al., 2017).

SearchGUI and its partner tool PeptideShaker are commonly used in proteomics analysis. These two types of software can be run on all of the three common computing

platforms, Windows, Mac, and Linux (Farag et al., 2021; Kopczynski et al., 2017). One strength of SearchGUI software is that SearchGUI supports more than 10 search engines, and SearchGUI results can be loaded directly to PeptideShaker without additional parameters. Both tools have a user-friendly graphical interface and an extensive command line interface. A command line interface should be used in high-performance computing (HPC) systems for large proteomics data analysis. Search time depends on the size of proteomics data and is also related to the search database. Search time increases as the size of proteomics data or search database increases (Shteynberg et al., 2013). In every MS search, we used 20 cores and ~160G memory. The size of MGF is, in total, 15-20G per organ/condition; the size of Fasta files is ~130,000 for altProt validation, ~80,000 for chimeric proteins validation modelled with MS-validated altProts and ~110,000 for chimeric proteins validation modelled with conserved altProts. Generally, the "barbun" partition was used to run MS searches in TRUBA. In this partition, RAM per Core is 8500M. However, the maximum size of the memory allocation pool of Java was set to 128G with the "Xmx" parameter because using the whole memory, ~160G, caused the incomplete protein reports or SearchGUI and/or PeptideShaker stopped responding. If the lower number of cores was set in SLURM job submission, a lower amount of memory can be used. With lower memory available, software may not finish the analysis properly. Besides, even though 20 cores with 128G memory provide relatively solid computational power, the running time of every MS search was ~24 hours, which indicates proteomic data analysis is computationally intensive and time-demanding. In altProt validation, 140 MS searches were conducted: 100 is first step for mRNA-derived, 10 is the second step for mRNA-derived, 30 is for non-mRNA-derived altProts. In chimeric proteins validation, 120 MS searches were conducted: 100 is the first step for chimeric proteins modelled with conserved altProts, 10 is the second step for chimeric proteins modelled with conserved altProts, and 10 is for chimeric proteins modelled with MS-validated altProts. The validation of altProts alone may take ~140 days if the searches are run in series, but we run them in parallel to decrease the total running time. In summary, 260 MS searches were conducted. We could run eight jobs in parallel; thus, the total running time was ~5 weeks.

4.7 Performance Analysis of the Two-Step Approach

On average, 9,930 (10-day nodules), 9,360 (14-day nodules), 5,930 (28-day nodules), 10,820 (buds), 10,920 (flowers), 8,640 (leaves), 6,220 (roots), 9,880 (seeds), 5,490 (stems), and 11,650 (whole plant) proteins were validated in the first step of the two-step approach. Standard deviation values of the identified proteins are as follows: 20 (10-day nodules), 1,060 (14-day nodules), 2,030 (28-day nodules), 30 (buds), 20 (flowers), 80 (leaves), 40 (roots), 10 (seeds), 90 (stems), and 30 (whole plant). Unexpectedly, the 4th group of the 28-day nodules sample has a very high FNR limit (97%) and a very low number of validated proteins (211). On the other hand, in the second step of the two-step approach, 10,020 (10-day nodules), 9,780 (14-day nodules), 2,610 (28-day nodules), 10,890 (buds), 13,030 (flowers), 8,640 (leaves), 6,130 (roots), 10,010 (seeds), 5,470 (stems), and 11,540 (whole plant) proteins were validated. Similar to the first step, the 28-day nodules sample has a very high FNR limit (61%) and a lower number of validated proteins (2,610) compared to other searches in the second step.

Non-mRNA-derived altProts were verified using the regular MS search approach. ncRNA, rRNA, and tRNA-derived altProts were searched separately using the same datasets, PXD002692 and PXD013606. Similar to mRNA-derived altProts, searches in the PXD002692 and PXD013606 datasets were performed independently.

In ncRNA-derived altProts MS searches, 9,920 (10-day nodules), 7,520 (14-day nodules), 370 (28-day nodules), 10,800 (buds), 12,900 (flowers), 8,570 (leaves), 6,200 (roots), 9,900 (seeds), 5,400 (stems), and 11,560 (whole plant) proteins were validated. Moreover, in MS searches for rRNA-derived altProts, 9,850 (10-day nodules), 9,730 (14-day nodules), 330 (28-day nodules), 10,770 (buds), 12,870 (flowers), 8,520 (leaves), 6,100 (roots), 9,860 (seeds), 5,410 (stems), and 11,450 (whole plant) proteins were validated. Furthermore, in tRNA-derived altProts MS searches, 9,880 (10-day nodules), 9,660 (14-day nodules), 340 (28-day nodules), 10,770 (buds), 12,860 (flowers), 8,480 (leaves), 6,120 (roots), 9,860 (seeds), 5,580 (stems), and 11,470 (whole plant) proteins were validated. Validation summary of all MS searches for altProts are shown in Table D.1-Table D.3.

Because we hypothesize that the number of translated canonical proteins, refProts, is much higher than translated altProts, the number of validated proteins should be approximately similar within conditions and organs. In other words, every search database in all searches has the same refProts plus different sets of altProts. For instance, ncRNA-derived altProts search in buds has a search database containing all refProts plus ncRNA-derived altProts, and rRNA-derived altProts search in buds has a search database containing all refProts plus rRNA-derived altProts. Because all refProts are common between them and the number of translated altProts is much less than refProts, at least it was assumed, the number of validated proteins in each search within organs/conditions should be the same if there are no interfering effects on MS searches such as non-randomized separation of the large database into equal groups for the first step of the two-step approach. In all searches, which include the first and second steps of the two-step approach for mRNA-derived altProts and searches for non-mRNA-derived altProts, the number of validated proteins within organs/conditions should be similar. The numbers of validated proteins from the first step of the two-step approach for mRNA-derived altProts are visualized in Figure 4.3, and the numbers of validated proteins from the second step of two-step approach for mRNA-derived altProts and non-mRNA-derived altProts searches are visualized in Figure 4.4. Approximately a similar number of proteins were validated within organs/conditions, with some exceptions. According to Figure 4.3, 14-day nodules and 28-day nodules have clearly different numbers of validated proteins. In the first step of the approach, the 4^{th} and 6^{th} groups of 14-day nodule-derived sequences have a relatively low number of validated proteins compared to other groups. Similarly, the 4^{th} group of 28-day nodules has a low number of validated proteins compared to other groups. According to Figure 4.4, in the second step of the two-step approach and in non-mRNA searches, the ncRNA-derived altProt search in 14-day nodules and the mRNA-derived altProt search in the 28-day nodules sample has a number of validated proteins different from others in their own group. Numerically, in the first step of the approach, Figure 4.3, the 4^{th} and 6^{th} groups have ~6,000 and ~9,000 validated proteins in the 14-day nodules search, and other groups in the same group have ~10,000 validated proteins. Furthermore, in the second step of the approach and non-mRNA-derived altProt searches, Figure 4.4, while ~10,000 proteins were validated in mRNA, rRNA, and tRNA searches, ~7,500 proteins were validated in ncRNA search in the 14-day nodules analysis. On the other hand, although less than 500 proteins were validated in ncRNA, rRNA, and tRNA searches, ~2,500 proteins were

validated in mRNA search in the 28-day nodules analysis. A comparison of the numbers of validated proteins indicates that our two-step approach is suitable for the inflated search databases since, most of the time, the number of validated proteins does not change drastically within an organ or condition. Overall, it indicates that the procedure may be adopted for studies on more complex MS proteomic datasets such as those available for humans.

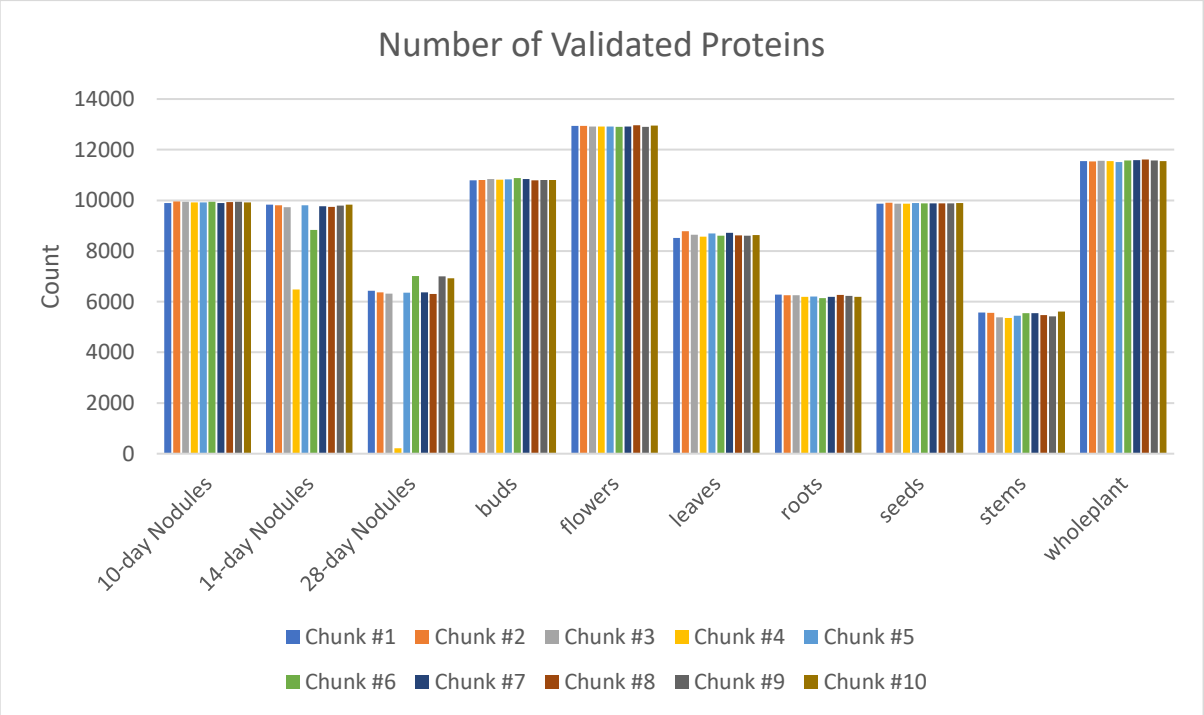

Figure 4.3. Number of validated proteins per organ/condition in mRNA-derived altProt searches. The data are shown for the first step of the two-step approach.

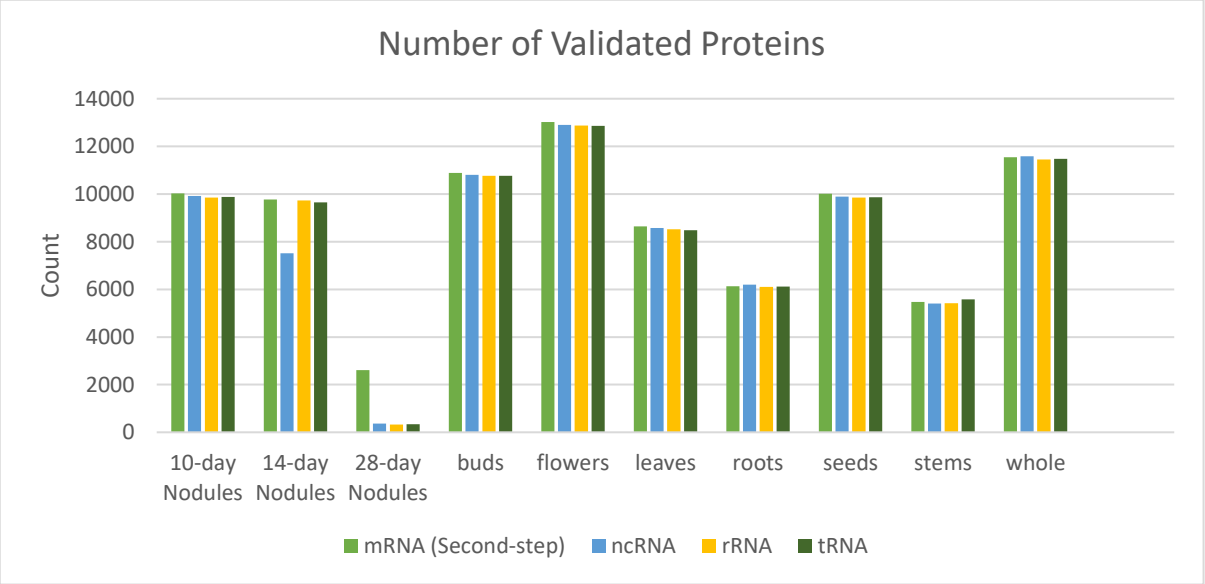

Figure 4.4. Number of validated proteins per organ/condition in different groups of transcripts. Numbers for mRNA-derived altProts are shown from the second step of the two-step approach.

4.8 What Do We Learn from Conservation Signatures of AltProts?

In this study, we define a conservation signature of an altProt as the presence of significant similarity (e-value 0.001 at most) to any protein annotated in the global database UniProt. This conservation signature has several parameters among which the most essential ones are % identity, the number of hits, the organism from which the top-scoring hit comes, and the annotation of the top hit subject. Based on these parameters, we can discriminate between intra- and inter-species conservation signatures. Intra-species conservation refers to the presence of one or several sequences similar to the altProt within the known proteome of *M. truncatula*. AltProts that have a single hit with 100% identity to a protein in *M. truncatula* are in the ambiguity group. These sequences may correspond either to overlapping genes or to transcribed genomic regions copied and inserted from other loci. Discrimination between these two cases requires dedicated analysis. To address this ambiguity, we conducted a preliminary analysis on a subset of nine ncRNA-derived altProts with 100% identity. Only one of such altProts corresponded to an overlapping protein-coding gene, and two other sequences were known as short peptides. Thus, the

remaining six altProts probably evolved from recent and ancient translocation events (Table G.1).

Inter-species conservation of an altProt corresponds to similarity with proteins from other organisms. Whereas the number of conserved altProts decreases with the evolutionary distance of the subject source species from *M. truncatula*, our study identified many altProts with top hits in non-plant organisms, including animals, fungi, protistans, and even prokaryotes. These altProts cannot be considered as candidates for recent trans-kingdom horizontal gene transfer events because some lower-scoring hits of these sequences come from legume plants (results not shown). Nevertheless, such altProts deserve close attention because they may have originated by trans-kingdom horizontal gene transfer events that predate the separation of *M. truncatula* from the common ancestor of legume plants.

Classically, the degree of conservation is thought to correlate with the functional importance of a protein sequence. Thus, conservation of an altProt at the protein level is strong evidence for translation. However, it should be noted that non-conserved proteins may have crucial roles in biological processes. Such *de novo* evolved sequences may quickly acquire functionality if they are integrated into the existing regulation networks (Schlötterer, 2015). Thus, altProts with no conservation signature but MS-based evidence for translation may be interesting targets for functional analysis.

Another important question relevant to this study is whether conserved sequences are translated. It is possible that mRNA- and ncRNA altProts with strong conservation signatures are indeed translated but do not appear in the MS proteomic datasets due to technical reasons. If so, does this also apply to rRNA and tRNA? Our data indicate that rRNA is associated with the highest occurrence of conserved altProts. Namely, ~ 47% of rRNA transcripts contained at least one conserved altProt (Table 3.2). This category was followed by tRNA, with ~ 32%. Remarkably, only ~ 6% of ncRNA sequences have at least one conserved altProt, which makes a large difference indicating a biological reason. Furthermore, the median % identity for tRNA was the highest among all groups, ~ 97%, followed by rRNA, with ~ 92%. What is so special about these two transcript groups? Does it mean that altORFs located in these transcripts are indeed translated? Translation of

functional polypeptides from rRNA has been documented (Root-Bernstein & Root-Bernstein, 2016). However, it is thought to be rather uncommon (only six rRNA-derived proteins are known so far). Our data indicate the opposite, even though no MS-supported peptide corresponding to rRNA was validated in our study. Concerning tRNA, translation from such molecules has never been detected and can hardly be anticipated because of their very short length. Nevertheless, we speculate that the detection of tRNA-encoded polypeptides is only a matter of time for two reasons. First, other very short species of RNA, namely pri-miRNA (Lauressergues et al., 2015; Sharma et al., 2019) and pri-siRNA (Yoshikawa et al., 2016), have been shown to encode polypeptides. Second, upon transcription, tRNA molecules are sent to the cytoplasm, where they are expected to be subjected to a phenomenon known as pervasive translation, like all other transcripts in the cell (Ingolia et al., 2014). There is little doubt that tRNA is associated with ribosomes. It is a matter of how functional that association can be. If the majority of rRNA and tRNA molecules are not translated, what is the biological purpose of maintaining this extraordinary degree of conservation at the protein level in these sequences? This is an exciting question with no answer so far. Possibly, this conservation reflects the central role of tRNA and rRNA in the evolution of genomes. Previously, it has been suggested that tRNAs are the proto-genes, the building blocks of the rRNA proto-genome (De Farias et al., 2016). According to this hypothesis, rRNA was formed by the polymerization of tRNA molecules that correspond to all the 20 usual proteinogenic amino acids. Initially, rRNA proto-genomes had a very high density of protein-coding, which is still reflected in their modern sequences. Our study fully supports this statement. Modern DNA genomes evolved from rRNA proto-genomes, in which individual proto-genes were gradually separated by non-coding sequences to facilitate the control of transcription. Nevertheless, it is inconceivable why the conservation in these transcripts is evident at the protein level instead of the nucleotide level. This may indicate that annotated proteins corresponding to top hits of these transcripts remained almost unchanged since the time when the tRNA-composed rRNA proto-genome was their only source. It also suggests that these proteins may still be produced from rRNA and possibly tRNA along with their homologs from mRNA. In this case, their functions may be different depending on the transcript type in cases where the top-hit % identity is below 100. Thus, at least theoretically, rRNA and tRNA should be considered as a potential source of proteins with unique essential functions.

4.9 Steps Towards Functional Characterization of AltProts, Chimeric Proteins, and Mosaic Proteins

AltProts, chimeric proteins, and mosaic proteins identified in our study constitute a resource of immense value for plant biologists who use *M. truncatula* to study symbiotic nitrogen fixation, symbiosis with arbuscular mycorrhizal fungi, and other fundamental biological processes. Some of these non-canonical proteins are associated with organ-specific transcripts. These should be subjected to in-depth functional analysis as they are the most likely candidates for specific roles in cellular processes. The methodology for elucidating functions of altProts is somewhat different from the standard set of procedures necessary to characterize conventional proteins. Here, we will outline an efficient protocol and will point out steps specific to studies on altProts.

First, for each candidate altProt, we have to learn which cell type, tissue, organ, or experimental condition is associated with the maximal expression of the corresponding transcript. Fortunately, this task is greatly facilitated by the availability of two major transcriptomic resources for *M. truncatula*, namely the RNA-Seq-based expression atlas MtExpress v2 (Carrere et al., 2021) and the Affymetrix microarray-based expression atlas MtGEA v3 (Benedito et al., 2008). Both resources are available via a user-friendly interface at https://medicago.toulouse.inrae.fr/GEA. In the next step, expression profiles obtained with the expression atlas must be confirmed using qRT-PCR and promoter-GUS fusion studies. In addition to the transcript location and the magnitude of transcription, the subcellular location of the candidate altProt must be determined using constructs in which the altProt is translationally fused to a fluorescent tag or a commonly used antibody epitope. Together with transcription profiling, this information is crucial for the next step, which is the validation of altProt translation using antibodies.

To raise antibodies specific to the altProt, a synthetic version of the altProt must be obtained via a commercial provider. Because we typically do not know the exact start and end position of the actually translated altProt, an epitope corresponding to the MS-validated peptide must be selected for antibody production. For chimeric proteins and mosaic proteins, the exact start and end of translation are also unknown. Due to the principal difference from individual translated altProts, chimeric proteins and mosaic

proteins require antibodies raised using special epitopes: sites corresponding to ribosomal frameshifts. Using individual parts as epitopes would be insufficient to demonstrate the continuity of the protein because a signal could correspond to altProts translated individually (without the fusion) from one transcript. Furthermore, multiple antibodies must be produced per mosaic protein to prove the continuity of the molecule. In this respect, colocalization of antibodies that correspond to frame-fusion sites of a single mosaic protein is essential. Thus, the secondary antibodies for their detection in biological samples must contain tags of different colours. These highly sensitive experiments must provide the desired resolution of subcellular localization along with the main goal, which is proving the fact of translation. In addition to these fine experiments, Western Blotting can be used as a complementing approach even though it can provide less information about the site of cellular activity of the candidate protein.

Knowing the exact start and end position of each non-canonical protein would give an advantage in deducing the entire amino acid sequence. Once the sequence is known in its entirety, many prediction tools can be applied to further help in its characterization: the prediction of topology, subcellular location, cleavage sites, secondary structures, folding, and even 3D structure, an option that became available only recently using artificial intelligence (Jumper et al., 2021). One of the greatest benefits of learning the start and end position of translation for such proteins is that antibodies can be raised specifically to the sequence as a whole, not to its individual parts. The need for this type of evidence is unique to chimeric proteins and mosaic proteins; translation of conventional proteins can be proved even with an MS-derived peptide partially covering its sequence. Knowing the entire sequence is also important for gain-of-function experiments, which should include ectopic expression of constructs designed to produce altProt, chimeric protein, and mosaic protein sequences without ribosomal frameshifting, using genetic frameshifting instead. Treatment of cells, tissues, organs, or entire organisms with synthetic proteins also is possible only if the sequence is known without any gap or truncation. Procedures that help detect the exact start and end of translation are too complex and too diverse to be covered in this thesis. However, they are routinely used in studies on altORFs in other organisms (Gagnon et al., 2021; Naville & Merabet, 2021; Vanderperre et al., 2013). Unfortunately, if translation of an altProt starts within the refProt of the same transcript, which is often the case, conventional methods such as RIBO-Seq are of little help in this respect.

4.10 Differential Mutagenesis of Overlapping ORFs is a Challenge

Loss-of-function studies are a necessary component of any comprehensive functional analysis. Often information that comes from such studies is the only evidence accepted as final proof of a biological role. Inferences based on gain-of-function experiments alone are less trustable and require many alternative lines of evidence to support them. The major challenge associated with the functional characterization of altProts is the need to differentially disable either the altORF or its overlapping refORF. Theoretically, this goal could be achieved using CRISPR/Cas9 technology (Adli, 2018). For example, a stop codon could be introduced in the middle of each sequence, thus resulting in two different mutant lines: one with truncated refORF but intact altORF and the other one with the opposite arrangement. However, taking into account the lack of 100% precision of CRISPR/Cas9 (Kelly et al., 2021), this approach is not very practical.

In the absence of a high-precision mutagenesis tool, what strategy can be the most informative for the differential mutagenesis of overlapping ORFs? Here, we propose a protocol that is indirect and requires more work than routine loss-of-function studies on refORFs. Nevertheless, it is probably the only currently possible way to characterize such proteins. In the first step, we need to determine if a simultaneous knockout of both the altORF and its refORF results in an altered phenotype. If it does not, neither the altORF nor its refORF is essential for the biological process under study. If the knockout is manifested in an observable disorder, there are three possibilities: (1) both the altORF and the refORF contribute to the phenotype; (2) only the altORF is essential; (3) only the refORF is essential. These generic mutants with both ORFs rendered non-functional are the basis for subsequent complementation studies, which are the key feature of this approach. To create such generic knockout transformants, insertional mutagenesis using retrotransposon *Tnt1* may be a method of choice because the mutants have already been established. They only need to be detected by PCR or by a database search (Lee et al., 2018). Deletion mutants are conceptually very useful for this purpose too. However, screening for deletion mutants is much more challenging (Williams et al., 2007). Chemical mutagenesis methods mostly produce populations of single-nucleotide mutants (Henikoff & Comai, 2003). This approach may seem to be much more useful for studying altORF because mutant lines can be found that are differentially affected in the altORF and its

refORF. Still, because the density of such mutations is very low, it is almost impossible to find two mutants containing nucleotide changes at desired positions. RNAi-based approaches are conceptually of little use for studying altORFs because they result in the constitutive degradation of the whole transcript, which makes it impossible to use complementation constructs in the background of RNAi transgenics (Travella & Keller, 2009). As was mentioned above, the ability to complement a generic mutant is crucial for the functional characterization of overlapping ORFs.

Thus, among the diverse mutagenesis methods available for *M. truncatula*, we recommend only one for studying altORFs: insertional mutagenesis using retrotransposon *Tnt1*. The disruption of genes using this method may occur in several ways. For example, the insertion may disable the promoter, which is an ideal outcome for disabling the whole gene so that no transcript is produced. If the insertion is in a coding sequence, there are several scenarios: (1) the entire transcript can be degraded; (2) the transcript can be partially degraded (truncated after the insertion); (3) the transcript can undergo major rearrangements that are mediated by abnormal splicing. This last scenario can involve the skipping of exons, retention of introns, and incorporation of the entire transposon or its parts into the aberrant transcript (J. Chen et al., 2006; Kryvoruchko et al., 2016; Menssen et al., 1990; Varagona et al., 1992). In all such cases, the refORF becomes entirely disabled while the overlapping altORF may either be disabled or retained without changes. The status of the affected altORF in such mutants can be tested by a regular PCR on cDNA from the mutant. If the altORF remains intact, the mutant gives information about the function of the refORF separately from the altORF. Further mutant lines can be identified among which at least one will have the altORF disabled together with the refORF.

The key feature of the approach we propose here is the complementation of such generic mutant lines using three different constructs, one containing the normal cDNA (or better gDNA) of the gene and the two other constructs containing either the refORF or the altORF disabled. Provided the fate of the transcript in the mutant background has been verified by PCR, the complementation phenotypes should give information on the contribution of the refORF, and separately of the altORF, into the trait. It is also possible that they have individual independent functions. In such a case, each complemented line will have a unique phenotype. If these complementation constructs are driven by a strong

constitutive promoter, they can also be used for gain-of-function studies for expression in the wild-type background. A clear idea about the function of each ORF in the overlapping pair should be formed based on the combination of all experimental domains mentioned above.

5. CONCLUSION

AltORFs and their translational products, altProts, have been shown to carry out essential functions in various organisms. We hypothesized that altORFs may act as building blocks for chimeric proteins and mosaic proteins, which are produced via single and multiple ribosomal frameshifting events from a mature transcript. Based on conservation analysis and MS searches, we validated 715 altProts and 146 ribosomal frameshifting positions that can support translation of chimeric proteins. Two transcripts are associated with more than two ribosomal frameshifting sites, which supports the existence of mosaic proteins and mosaic translation. We validated two mosaic proteins which have never been detected in non-viral organisms before. This dataset is so far unique and will be of high interest not only for plant biologists but also for researchers from other fields. If frameshifted proteins are validated by follow-up wet-bench experiments, it pioneers a new field of proteomic studies and paves the road towards the discovery of non-viral proteins of chimeric and mosaic nature in higher eukaryotes, including humans.

6. REFERENCES

Adli, M., “The CRISPR Tool Kit for Genome Editing and Beyond”, *Nature Communications*, Vol. 9, No. 1, pp. 1–13, Nature Publishing Group, 2018.

Adusumilli, R., and P. Mallick, “Data Conversion with ProteoWizard MsConvert”, *Methods in Molecular Biology*, Vol. 1550, No. 1, pp. 339–368, 2017.

Advani, V. M., and J. D. Dinman, “Reprogramming the Genetic Code: The Emerging Role of Ribosomal Frameshifting in Regulating Cellular Gene Expression”, *BioEssays*, Vol. 38, No. 1, pp. 21–26, 2016.

Anders, J., H. Petruschke, N. Jehmlich, S. B. Haange, M. von Bergen and P. F. Stadler, “A Workflow to Identify Novel Proteins Based on the Direct Mapping of Peptide-Spectrum-Matches to Genomic Locations”, *BMC Bioinformatics*, Vol. 22, No. 1, pp. 1–20, 2021.

Andreev, D. E., G. Loughran, A. D. Fedorova, M. S. Mikhaylova, I. N. Shatsky and P. V. Baranov, “Non-AUG Translation Initiation in Mammals”, *Genome Biology*, Vol. 23, No. 1, pp. 1–17, 2022.

Ané, J. M., H. Zhu and J. Frugoli, “Recent Advances in Medicago Truncatula Genomics”, *International Journal of Plant Genomics*, Vol. 2008, No. 1, p. 11, 2008.

Aspden, J. L., Y. C. Eyre-Walker, R. J. Phillips, U. Amin, M. A. S. Mumtaz, M. Brocard and J. P. Couso, “Extensive Translation of Small Open Reading Frames Revealed by Poly-Ribo-Seq”, *ELife*, Vol. 3, No. August 2014, pp. 1–19, 2014.

Atkins, J. F., G. Loughran, P. R. Bhatt, A. E. Firth and P. V. Baranov, “Ribosomal Frameshifting and Transcriptional Slippage: From Genetic Steganography and Cryptography to Adventitious Use”, *Nucleic Acids Research*, Vol. 44, No. 15, pp. 7007–7078, 2016.

Baker, M. S., S. B. Ahn, A. Mohamedali, M. T. Islam, D. Cantor, P. D. Verhaert, S. Fanayan, S. Sharma, E. C. Nice, M. Connor and S. Ranganathan, "Accelerating the Search for the Missing Proteins in the Human Proteome", *Nature Communications*, Vol. 8, No. 1, pp. 1–13, 2017.

Barsnes, H., and M. Vaudel, "SearchGUI: A Highly Adaptable Common Interface for Proteomics Search and de Novo Engines", *Journal of Proteome Research*, Vol. 17, No. 7, pp. 2552–2555, 2018.

Benedito, V. A., I. Torres-Jerez, J. D. Murray, A. Andriankaja, S. Allen, K. Kakar, M. Wandrey, J. Verdier, H. Zuber, T. Ott, S. Moreau, A. Niebel, T. Frickey, G. Weiller, J. He, X. Dai, P. X. Zhao, Y. Tang and M. K. Udvardi, "A Gene Expression Atlas of the Model Legume Medicago Truncatula", *Plant Journal*, Vol. 55, No. 3, pp. 504–513, 2008.

Benitez-Cantos, M. S., M. M. Yordanova, P. B. F. O'Connor, A. V. Zhdanov, S. I. Kovalchuk, D. B. Papkovsky, D. E. Andreev and P. V. Baranov, "Translation Initiation Downstream from Annotated Start Codons in Human MRNAs Coevolves with the Kozak Context", *Genome Research*, Vol. 30, No. 7, pp. 974–984, 2020.

Beyter, D., M. S. Lin, Y. Yu, R. Pieper and V. Bafna, "ProteoStorm: An Ultrafast Metaproteomics Database Search Framework", *Cell Systems*, Vol. 7, No. 4, pp. 463-467.e6, 2018.

Bhatt, P. R., A. Scaiola, G. Loughran, M. Leibundgut, A. Kratzel, R. Meurs, R. Dreos, K. M. O'Connor, A. McMillan, J. W. Bode, V. Thiel, D. Gatfield, J. F. Atkins and N. Ban, "Structural Basis of Ribosomal Frameshifting during Translation of the SARS-CoV-2 RNA Genome", *Science*, Vol. 372, No. 6548, pp. 1306–1313, 2021.

Bischoff, L., O. Berninghausen and R. Beckmann, "Molecular Basis for the Ribosome Functioning as an L-Tryptophan Sensor", *Cell Reports*, Vol. 9, No. 2, pp. 469–475, 2014.

Biswas, P., X. Jiang, A. L. Pacchia, J. P. Dougherty and S. W. Peltz, "The Human

Immunodeficiency Virus Type 1 Ribosomal Frameshifting Site Is an Invariant Sequence Determinant and an Important Target for Antiviral Therapy", *Journal of Virology*, Vol. 78, No. 4, pp. 2082–2087, 2004.

Bludau, I., and R. Aebersold, "Proteomic and Interactomic Insights into the Molecular Basis of Cell Functional Diversity", *Nature Reviews Molecular Cell Biology*, Vol. 21, No. 6, pp. 327–340, 2020.

Borsovszky, J., K. Nauta, J. Jiang, C. S. Hansen, L. K. McKemmish, R. W. Field, J. F. Stanton, S. H. Kable and T. W. Schmidt, "Photodissociation of Dicarbon: How Nature Breaks an Unusual Multiple Bond", *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 118, No. 52, p. e2113315118, 2021.

Brierley, I., R. J. C. Gilbert and S. Pennell, "Pseudoknot-Dependent Programmed —1 Ribosomal Frameshifting: Structures, Mechanisms and Models", *Recoding: Expansion of Decoding Rules Enriches Gene Expression*, Vol. 24, No. 1, pp. 149–174, 2010.

Brunet, M. A., S. A. Levesque, D. J. Hunting, A. A. Cohen and X. Roucou, "Recognition of the Polycistronic Nature of Human Genes Is Critical to Understanding the Genotype-Phenotype Relationship", *Genome Research*, Vol. 28, No. 5, pp. 609–624, 2018.

Brunet, M. A., J. F. Lucier, M. Levesque, S. Leblanc, J. F. Jacques, H. R. H. Al-Saedi, N. Guilloy, F. Grenier, M. Avino, I. Fournier, M. Salzet, A. Ouangraoua, M. S. Scott, F. M. Boisvert and X. Roucou, "OpenProt 2021: Deeper Functional Annotation of the Coding Potential of Eukaryotic Genomes", *Nucleic Acids Research*, Vol. 49, No. D1, pp. D380–D388, 2021.

Buchfink, B., K. Reuter and H. G. Drost, "Sensitive Protein Alignments at Tree-of-Life Scale Using DIAMOND", *Nature Methods*, Vol. 18, No. 4, pp. 366–368, 2021.

Buchfink, B., C. Xie and D. H. Huson, "Fast and Sensitive Protein Alignment Using DIAMOND", *Nature Methods*, Vol. 12, No. 1, pp. 59–60, 2014.

Büyükköroğlu, G., D. D. Dora, F. Özdemir and C. Hızel, “Chapter 15 - Techniques for Protein Analysis”, D. Barh & V. Azevedo Eds., *Omics Technologies and Bio-Engineering*, pp. 317–351, Academic Press, 2018.

Byun, Y., S. Moon and K. Han, “Prediction of Ribosomal Frameshift Signals of User-Defined Models”, *Lecture Notes in Computer Science*, Vol. 3514, No. I, pp. 948–955, 2005.

Çakır, U., N. Gabed, M. Brunet, X. Roucou and I. Kryvoruchko, “Mosaic Translation Hypothesis: Chimeric Polypeptides Produced via Multiple Ribosomal Frameshifting as a Basis for Adaptability [Published Online Ahead of Print, 2021 Nov 7]”, *FEBS Journal*, John Wiley & Sons, Ltd, 2021.

Caliskan, N., V. I. Katunin, R. Belardinelli, F. Peske and M. V. Rodnina, “Programmed -1 Frameshifting by Kinetic Partitioning during Impeded Translocation”, *Cell*, Vol. 157, No. 7, pp. 1619–1631, 2014.

Caliskan, N., I. Wohlgemuth, N. Korniy, M. Pearson, F. Peske and M. V. Rodnina, “Conditional Switch between Frameshifting Regimes upon Translation of DnaX MRNA”, *Molecular Cell*, Vol. 66, No. 4, pp. 558-567.e4, 2017.

Camacho, C., G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer and T. L. Madden, “BLAST+: Architecture and Applications”, *BMC Bioinformatics*, Vol. 10, No. 1, p. 421, 2009.

Cao, X., and S. A. Slavoff, “Non-AUG Start Codons: Expanding and Regulating the Small and Alternative ORFeome”, *Experimental Cell Research*, Vol. 391, No. 1, p. 111973, 2020.

Cao, Y., and L. Ma, “To Splice or to Transcribe: SKIP-Mediated Environmental Fitness and Development in Plants”, *Frontiers in Plant Science*, Vol. 10, No. 1, p. 1222, 2019.

Cardon, T., J. Franck, E. Coyaud, E. M. N. Laurent, M. Damato, M. Maffia, D. Vergara, I.

Fournier and M. Salzet, “Alternative Proteins Are Functional Regulators in Cell Reprogramming by PKA Activation”, *Nucleic Acids Research*, Vol. 48, No. 14, pp. 7864–7882, 2020.

Carrere, S., J. Verdier and P. Gamas, “MtExpress, a Comprehensive and Curated Rnaseq-Based Gene Expression Atlas for the Model Legume Medicago Truncatula”, *Plant and Cell Physiology*, Vol. 62, No. 9, pp. 1494–1500, 2021.

Charon, C., C. Johansson, E. Kondorosi, A. Kondorosi and M. Crespi, “Enod40 Induces Dedifferentiation and Division of Root Cortical Cells in Legumes”, *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 94, No. 16, pp. 8901–8906, 1997.

Charon, J., S. Theil, V. Nicaise and T. Michon, “Protein Intrinsic Disorder within the Potyvirus Genus: From Proteome-Wide Analysis to Functional Annotation”, *Molecular BioSystems*, Vol. 12, No. 2, pp. 634–652, 2016.

Chatterjee, S., G. S. Stupp, S. K. R. Park, J. C. Ducom, J. R. Yates, A. I. Su and D. W. Wolan, “A Comprehensive and Scalable Database Search System for Metaproteomics”, *BMC Genomics*, Vol. 17, No. 1, pp. 1–11, 2016.

Cheetham, S. W., G. J. Faulkner and M. E. Dinger, “Overcoming Challenges and Dogmas to Understand the Functions of Pseudogenes”, *Nature Reviews Genetics*, Vol. 21, No. 3, pp. 191–201, 2020.

Chen, C., J. Hou, J. J. Tanner and J. Cheng, “Bioinformatics Methods for Mass Spectrometry-Based Proteomics Data Analysis”, *International Journal of Molecular Sciences*, Vol. 21, No. 8, p. 2873, 2020.

Chen, J., A. Rattner and J. Nathans, “Effects of L1 Retrotransposon Insertion on Transcript Processing, Localization and Accumulation: Lessons from the Retinal Degeneration 7 Mouse and Implications for the Genomic Ecology of L1 Elements”, *Human Molecular Genetics*, Vol. 15, No. 13, pp. 2146–2156, 2006.

Chen, X., H. Kang, L. X. Shen, M. Chamorro, H. E. Varmus and I. Tinoco, “A Characteristic Bent Conformation of RNA Pseudoknots Promotes - 1. Frameshifting during Translation of Retroviral RNA”, *Journal of Molecular Biology*, Vol. 260, No. 4, pp. 479–483, 1996.

Cheng-Guang, H., and C. O. Gualerzi, “The Ribosome as a Switchboard for Bacterial Stress Response”, *Frontiers in Microbiology*, Vol. 11, No. 1, p. 619038, 2021.

Cheng, X., H. Xie, K. Zhang and J. Wen, “Enabling Medicago Truncatula Forward Genetics: Identification of Genetic Crossing Partner for R108 and Development of Mapping Resources for Tnt1 Mutants”, *The Plant Journal*, Vol. 111, No. 2, pp. 608–616, 2022.

Choi, J., Z. Xu and J. Ou, “Triple Decoding of Hepatitis C Virus RNA by Programmed Translational Frameshifting”, *Molecular and Cellular Biology*, Vol. 23, No. 5, pp. 1489–1497, 2003.

Chorev, D. S., G. Ben-Nissan and M. Sharon, “Exposing the Subunit Diversity and Modularity of Protein Complexes by Structural Mass Spectrometry Approaches”, *Proteomics*, Vol. 15, No. 16, pp. 2777–2791, 2015.

Claverie, J. M., O. Poirot and F. Lopez, “The Difficulty of Identifying Genes in Anonymous Vertebrate Sequences*”, *Computers and Chemistry*, Vol. 21, No. 4, pp. 203–214, 1997.

Cook, D. R., “Medicago Truncatula - A Model in the Making!”, *Current Opinion in Plant Biology*, Vol. 2, No. 4, pp. 301–304, 1999.

Curtin, S. J., P. Tiffin, J. Guhlin, D. Trujillo, L. Burghart, P. Atkins, N. J. Baltes, R. Denny, D. F. Voytas, R. M. Stupar and N. D. Young, “Validating Genome-Wide Association Candidates Controlling Quantitative Variation in Nodulation”, *Plant Physiology*, Vol. 173, No. 2, pp. 921–931, 2017.

De Bruijn, F. J., “Symbiotic Nitrogen Fixation”, *The Model Legume Medicago truncatula*,

pp. 429–431, John Wiley & Sons, Ltd, 2019.

De Farias, S. T., T. G. Rêgo and M. V. José, “TRNA Core Hypothesis for the Transition from the RNA World to the Ribonucleoprotein World”, *Life*, Vol. 6, No. 2, p. 15, 2016.

Deonier, R. C., M. S. Waterman and S. Tavaré, “Computational Genome Analysis”, *Computational Genome Analysis*, Springer New York, 2005.

Deutsch, E. W., “File Formats Commonly Used in Mass Spectrometry Proteomics”, *Molecular and Cellular Proteomics*, Vol. 11, No. 12, pp. 1612–1621, 2012.

Deutsch, E. W., N. Bandeira, V. Sharma, Y. Perez-Riverol, J. J. Carver, D. J. Kundu, D. García-Seisdedos, A. F. Jarnuczak, S. Hewapathirana, B. S. Pullman, J. Wertz, Z. Sun, S. Kawano, S. Okuda, Y. Watanabe, H. Hermjakob, B. Maclean, M. J. Maccoss, Y. Zhu, Y. Ishihama and J. A. Vizcaíno, “The ProteomeXchange Consortium in 2020: Enabling “big Data” Approaches in Proteomics”, *Nucleic Acids Research*, Vol. 48, No. D1, pp. D1145–D1152, 2020.

Dinman, J. D., “Programmed Ribosomal Frameshifting Goes beyond Viruses”, *Microbe*, Vol. 1, No. 11, pp. 521–527, Microbe Wash DC, 2006.

Dinman, J. D., “Pathways to Specialized Ribosomes: The Brussels Lecture”, *Journal of Molecular Biology*, Vol. 428, No. 10, pp. 2186–2194, 2016.

Dinman, J. D., T. Icho and R. B. Wickner, “A -1 Ribosomal Frameshift in a Double-Stranded RNA Virus of Yeast Forms a Gag-Pol Fusion Protein”, *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 88, No. 1, pp. 174–178, 1991.

Dinman, J. D., S. Richter, E. P. Plant, R. C. Taylor, A. B. Hammell and T. M. Rana, “The Frameshift Signal of HIV-1 Involves a Potential Intramolecular Triplex RNA Structure”, *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 99, No. 8, pp. 5331–5336, 2002.

Doane, T. A., “The Abiotic Nitrogen Cycle”, *ACS Earth and Space Chemistry*, Vol. 1, No. 7, pp. 411–421, 2017.

Domonkos, A., B. Horvath, J. F. Marsh, G. Halasz, F. Ayaydin, G. E. D. Oldroyd and P. Kalo, “The Identification of Novel Loci Required for Appropriate Nodule Development in Medicago Truncatula”, *BMC Plant Biology*, Vol. 13, No. 1, pp. 1–11, 2013.

Dong, W., Y. Zhu, H. Chang, C. Wang, J. Yang, J. Shi, J. Gao, W. Yang, L. Lan, Y. Wang, X. Zhang, H. Dai, Y. Miao, L. Xu, Z. He, C. Song, S. Wu, D. Wang, N. Yu and E. Wang, “An SHR–SCR Module Specifies Legume Cortical Cell Fate to Enable Nodulation”, *Nature*, Vol. 589, No. 7843, pp. 586–590, 2021.

Eckardt, N. A., “The Role of Flavonoids in Root Nodule Development and Auxin Transport in Medicago Truncatula”, *Plant Cell*, Vol. 18, No. 7, pp. 1539–1540, 2006.

Esseling, J. J., F. G. P. Lhuissier and A. M. C. Emons, “Nod Factor-Induced Root Hair Curling: Continuous Polar Growth towards the Point of Nod Factor Application”, *Plant Physiology*, Vol. 132, No. 4, pp. 1982–1988, 2003.

Fabre, B., S. A. Choteau, C. Duboé, C. Pichereaux, A. Montigny, D. Korona, M. J. Deery, M. Camus, C. Brun, O. Burlet-Schiltz, S. Russell, J. P. Combier, K. S. Lilley and S. Plaza, “In Depth Exploration of the Alternative Proteome of Drosophila Melanogaster”, *Frontiers in Cell and Developmental Biology*, Vol. 10, No. 1, p. 901351, 2022.

Farabaugh, P., “Translational Frameshifting, Non-Standard Reading of the Genetic Code”, *Encyclopedic Reference of Genomics and Proteomics in Molecular Medicine*, pp. 1910–1913, Springer, Berlin, Heidelberg, 2006.

Farag, Y. M., C. Horro, M. Vaudel and H. Barsnes, “PeptideShaker Online: A User-Friendly Web-Based Framework for the Identification of Mass Spectrometry-Based Proteomics Data”, *Journal of Proteome Research*, Vol. 20, No. 12, pp. 5419–5423, 2021.

Farajollahi, S., and S. Maas, “Molecular Diversity through RNA Editing: A Balancing Act”, *Trends in Genetics*, Vol. 26, No. 5, pp. 221–230, 2010.

Farkas, A., G. Maróti, H. Dürgo, Z. Györgypál, R. M. Lima, K. F. Medzihradszky, A. Kereszt, P. Mergaert and É. Kondorosi, “Medicago Truncatula Symbiotic Peptide NCR247 Contributes to Bacteroid Differentiation through Multiple Mechanisms”, *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 111, No. 14, pp. 5183–5188, 2014.

Ferretti, M. B., and K. Karbstein, “Does Functional Specialization of Ribosomes Really Exist?”, *Rna*, Vol. 25, No. 5, pp. 521–538, 2019.

Filip, S., K. Vougas, J. Zoidakis, A. Latosinska, W. Mullen, G. Spasovski, H. Mischak, A. Vlahou and J. Jankowski, “Comparison of Depletion Strategies for the Enrichment of Low-Abundance Proteins in Urine”, *PLoS ONE*, Vol. 10, No. 7, p. e0133773, 2015.

Firth, A. E., B. W. Jagger, H. M. Wise, C. C. Nelson, K. Parsawar, N. M. Wills, S. Napthine, J. K. Taubenberger, P. Digard and J. F. Atkins, “Ribosomal Frameshifting Used in Influenza A Virus Expression Occurs within the Sequence UCC-UUU-CGU and Is in the +1 Direction”, *Open Biology*, Vol. 2, No. 10, p. 120109, 2012.

Frendo, P., J. Harrison, C. Norman, M. J. H. Jiménez, G. Van De Sype, A. Gilabert and A. Puppo, “Glutathione and Homoglutathione Play a Critical Role in the Nodulation Process of Medicago Truncatula”, *Molecular Plant-Microbe Interactions*, Vol. 18, No. 3, pp. 254–259, 2005.

Frendo, P., M. J. Hernández Jiménez, C. Mathieu, L. Duret, D. Gallesi, G. Van de Sype, D. Hérouart and A. Puppo, “A Medicago Truncatula Homoglutathione Synthetase Is Derived from Glutathione Synthetase by Gene Duplication”, *Plant Physiology*, Vol. 126, No. 4, pp. 1706–1715, 2001.

Fu, Q., and L. Li, “De Novo Sequencing of Neuropeptides Using Reductive Isotopic Methylation and Investigation of ESI QTOF MS/MS Fragmentation Pattern of Neuropeptides with N-Terminal Dimethylation”, *Analytical Chemistry*, Vol. 77, No.

23, pp. 7783–7795, 2005.

Gagnon, M., M. Savard, J. F. Jacques, G. Bkaily, S. Geha, X. Roucou and F. Gobeil, "Potentiation of B2 Receptor Signaling by AltB2R, a Newly Identified Alternative Protein Encoded in the Human Bradykinin B2 Receptor Gene", *Journal of Biological Chemistry*, Vol. 296, No. 1, p. 100329, 2021.

Gavrin, A., and S. Schornack, "Medicago Truncatula as a Model Organism to Study Conserved and Contrasting Aspects of Symbiotic and Pathogenic Signaling Pathways", *The Model Legume Medicago truncatula*, pp. 317–330, John Wiley & Sons, Ltd, 2019.

Gerstein, M. B., C. Bruce, J. S. Rozowsky, D. Zheng, J. Du, J. O. Korbel, O. Emanuelsson, Z. D. Zhang, S. Weissman and M. Snyder, "What Is a Gene, Post-ENCODE? History and Updated Definition", *Genome Research*, Vol. 17, No. 6, pp. 669–681, 2007.

Gilbert, W. V., "Functional Specialization of Ribosomes?", *Trends in Biochemical Sciences*, Vol. 36, No. 3, pp. 127–132, 2011.

Glish, G. L., and D. J. Burinsky, "Hybrid Mass Spectrometers for Tandem Mass Spectrometry", *Journal of the American Society for Mass Spectrometry*, Vol. 19, No. 2, pp. 161–172, 2008.

Gregorich, Z. R., Y. H. Chang and Y. Ge, "Proteomics in Heart Failure: Top-down or Bottom-Up?", *Pflugers Archiv European Journal of Physiology*, Vol. 466, No. 6, pp. 1199–1209, 2014.

Guerra-Almeida, D., and R. Nunes-da-Fonseca, "Small Open Reading Frames: How Important Are They for Molecular Evolution?", *Frontiers in Genetics*, Vol. 11, No. 1, p. 574737, 2020.

Guerra-Almeida, D., D. A. Tschoeke and R. Nunes-Da-Fonseca, "Understanding Small ORF Diversity through a Comprehensive Transcription Feature Classification", *DNA Research*, Vol. 28, No. 5, pp. 1–18, 2021.

Guo, H., “Specialized Ribosomes and the Control of Translation”, *Biochemical Society Transactions*, Vol. 46, No. 4, pp. 855–869, 2018.

Gurvich, O. L., P. V. Baranov, R. F. Gesteland and J. F. Atkins, “Expression Levels Influence Ribosomal Frameshifting at the Tandem Rare Arginine Codons AGG_AGG and AGA_AGA in Escherichia Coli”, *Journal of Bacteriology*, Vol. 187, No. 12, pp. 4023–4032, 2005.

Han, X., A. Aslanian and J. R. Yates, “Mass Spectrometry for Proteomics”, *Current Opinion in Chemical Biology*, Vol. 12, No. 5, pp. 483–490, 2008.

Hansen, T. M., S. Nader S Reihani, L. B. Oddershede and M. A. Sørensen, “Correlation between Mechanical Strength of Messenger RNA Pseudoknots and Ribosomal Frameshifting”, *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 104, No. 14, pp. 5830–5835, 2007.

Hatfield, D. L., J. G. Levin, A. Rein and S. Oroszlan, “Translational Suppression in Retroviral Gene Expression”, *Advances in Virus Research*, Vol. 41, No. C, pp. 193–239, 1992.

He, L., J. Diedrich, Y. Y. Chu and J. R. Yates, “Extracting Accurate Precursor Information for Tandem Mass Spectra by RawConverter”, *Analytical Chemistry*, Vol. 87, No. 22, pp. 11361–11367, 2015.

He, L., and B. Ma, “Adepts: Advanced Peptide de Novo Sequencing with a Pair of Tandem Mass Spectra”, *Journal of Bioinformatics and Computational Biology*, Vol. 8, No. 6, pp. 981–994, 2010.

He, Z. shui, H. song Zou, Y. zhang Wang, J. bi Zhu and G. qiao Yu, “Maturation of the Nodule-Specific Transcript MsHSF1c in Medicago Sativa May Involve Interallelic Trans-Splicing”, *Genomics*, Vol. 92, No. 2, pp. 115–121, 2008.

Heather, J. M., and B. Chain, “The Sequence of Sequencers: The History of Sequencing DNA”, *Genomics*, Vol. 107, No. 1, pp. 1–8, 2016.

Henikoff, S., and L. Comai, “Single-Nucleotide Mutations for Plant Functional Genomics”, *Annual Review of Plant Biology*, Vol. 54, No. 1, pp. 375–401, 2003.

Horvath, B., A. Domonkos, A. Kereszt, A. Szucs, E. Abraham, F. Ayaydin, K. Boka, Y. Chen, R. Chen, J. D. Murray, M. K. Udvardi, E. Kondorosi and P. Kalo, “Loss of the Nodule-Specific Cysteine Rich Peptide, NCR169, Abolishes Symbiotic Nitrogen Fixation in the Medicago Truncatula Dnf7 Mutant”, *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 112, No. 49, pp. 15232–15237, 2015.

Hulstaert, N., J. Shofstahl, T. Sachsenberg, M. Walzer, H. Barsnes, L. Martens and Y. Perez-Riverol, “ThermoRawFileParser: Modular, Scalable, and Cross-Platform RAW File Conversion”, *Journal of Proteome Research*, Vol. 19, No. 1, pp. 537–542, 2020.

Ingolia, N. T., G. A. Brar, N. Stern-Ginossar, M. S. Harris, G. J. S. Talhouarne, S. E. Jackson, M. R. Wills and J. S. Weissman, “Ribosome Profiling Reveals Pervasive Translation Outside of Annotated Protein-Coding Genes”, *Cell Reports*, Vol. 8, No. 5, pp. 1365–1379, 2014.

Jacks, T., “Translational Suppression in Gene Expression in Retroviruses and Retrotransposons”, *Current Topics in Microbiology and Immunology*, Vol. 157, No. 1, pp. 93–124, 1990.

Jensen, K. T., L. Petersen, S. Falk, P. Iversen, P. Andersen, M. Theisen and A. Krogh, “Novel Overlapping Coding Sequences in Chlamydia Trachomatis”, *FEMS Microbiology Letters*, Vol. 265, No. 1, pp. 106–117, 2006.

Jensen, L. J., D. W. Ussery and S. Brunak, “Functionality of System Components: Conservation of Protein Function in Protein Feature Space”, *Genome Research*, Vol. 13, No. 11, pp. 2444–2449, 2003.

Johnson, R. S., and J. A. Taylor, “Searching Sequence Databases via de Novo Peptide Sequencing by Tandem Mass Spectrometry.”, *Methods in Molecular Biology (Clifton, N.J.)*, Vol. 146, No. 1, pp. 41–61, 2000.

Jumper, J., R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli and D. Hassabis, "Highly Accurate Protein Structure Prediction with AlphaFold", *Nature*, Vol. 596, No. 7873, pp. 583–589, 2021.

Kang, Y., M. Li, S. Sinharoy and J. Verdier, "A Snapshot of Functional Genetic Studies in Medicago Truncatula", *The Model Legume Medicago Truncatula*, Vol. 7, No. AUG2016, pp. 7–30, 2019.

Karaoglanoglu, F., C. Chauve and F. Hach, "Genion, an Accurate Tool to Detect Gene Fusion from Long Transcriptomics Reads", *BMC Genomics*, Vol. 23, No. 1, pp. 23–129, 2022.

Karginov, T. A., D. P. H. Pastor, B. L. Semler and C. M. Gomez, "Mammalian Polycistronic MRNAs and Disease", *Trends in Genetics*, Vol. 33, No. 2, pp. 129–142, NIH Public Access, 2017.

Kartali, T., I. Nyilasi, S. Kocsubé, R. Patai, T. F. Polgár, N. Zsindely, G. Nagy, L. Bodai, Z. Lipinszki, C. Vágvölgyi and T. Papp, "Characterization of Four Novel Dsrna Viruses Isolated from Mucor Hiemalis Strains", *Viruses*, Vol. 13, No. 11, p. 2319, 2021.

Kawakami, K., S. Pande, B. Faiola, D. P. Moore, J. D. Boeke, P. J. Farabaugh, J. N. Strathern, Y. Nakamura and D. J. Garfinkel, "A Rare TRNA-Arg(CCU) That Regulates Ty1 Element Ribosomal Frameshifting Is Essential for Ty1 Retrotransposition in Saccharomyces Cerevisiae", *Genetics*, Vol. 135, No. 2, pp. 309–320, 1993.

Kearse, M. G., and J. E. Wilusz, "Non-AUG Translation: A New Start for Protein Synthesis in Eukaryotes", *Genes and Development*, Vol. 31, No. 17, pp. 1717–1731, 2017.

Kelly, J. J., M. Saee-Marand, N. N. Nyström, M. M. Evans, Y. Chen, F. M. Martinez, A. M. Hamilton and J. A. Ronald, "Safe Harbor-Targeted CRISPR-Cas9 Homology-Independent Targeted Integration for Multimodality Reporter Gene-Based Cell Tracking", *Science Advances*, Vol. 7, No. 4, p. eabc3791, 2021.

Kertesz-Farkas, A., B. Reiz, M. P. Myers and S. Pongor, "Database Searching in Mass Spectrometry Based Proteomics", *Current Bioinformatics*, Vol. 7, No. 2, pp. 221–230, 2012.

Ketteler, R., "On Programmed Ribosomal Frameshifting: The Alternative Proteomes", *Frontiers in Genetics*, Vol. 3, No. NOV, p. 242, Frontiers Media SA, 2012.

Khitun, A., and S. A. Slavoff, "Proteomic Detection and Validation of Translated Small Open Reading Frames", *Current Protocols in Chemical Biology*, Vol. 11, No. 4, p. e77, 2019.

Kim, M., Y. Chen, J. Xi, C. Waters, R. Chen and D. Wang, "An Antimicrobial Peptide Essential for Bacterial Survival in the Nitrogen-Fixing Symbiosis", *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 112, No. 49, pp. 15238–15243, 2015.

Kim, S., and P. A. Pevzner, "MS-GF+ Makes Progress towards a Universal Database Search Tool for Proteomics", *Nature Communications*, Vol. 5, No. 1, pp. 1–10, 2014.

Kima, H. K., F. Liua, J. Fei, C. Bustamante, R. L. Gonzalez and I. Tinoco, "A Frameshifting Stimulatory Stem Loop Destabilizes the Hybrid State and Impedes Ribosomal Translocation", *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 111, No. 15, pp. 5538–5543, 2014.

Kollmus, H., A. Honigman, A. Panet and H. Hauser, "The Sequences of and Distance between Two Cis-Acting Signals Determine the Efficiency of Ribosomal Frameshifting in Human Immunodeficiency Virus Type 1 and Human T-Cell Leukemia Virus Type II in Vivo", *Journal of Virology*, Vol. 68, No. 9, pp. 6087–6091, 1994.

Kong, Y., Z. Meng, H. Wang, Y. Wang, Y. Zhang, L. Hong, R. Liu, M. Wang, J. Zhang, L. Han, M. Bai, X. Yu, F. Kong, K. S. Mysore, J. Wen, P. Xin, J. Chu and C. Zhou, "Brassinosteroid Homeostasis Is Critical for the Functionality of the Medicago Truncatula Pulvinus", *Plant Physiology*, Vol. 185, No. 4, pp. 1745–1763, 2021.

Kopczynski, D., A. Sickmann and R. Ahrends, "Computational Proteomics Tools for Identification and Quality Control", *Journal of Biotechnology*, Vol. 261, No. 1, pp. 126–130, 2017.

Korniy, N., E. Samatova, M. M. Anokhina, F. Peske and M. V. Rodnina, "Mechanisms and Biomedical Implications of –1 Programmed Ribosome Frameshifting on Viral and Bacterial MRNAs", *FEBS Letters*, Vol. 593, No. 13, pp. 1468–1482, 2019.

Kryvoruchko, I. S., P. Routray, S. Sinharoy, I. Torres-Jerez, M. Tejada-Jiménez, L. A. Finney, J. Nakashima, C. I. Pislariu, V. A. Benedito, M. González-Guerrero, D. M. Roberts and M. K. Udvardi, "An Iron-Activated Citrate Transporter, MtMATE67, Is Required for Symbiotic Nitrogen Fixation", *Plant Physiology*, Vol. 176, No. 3, pp. 2315–2329, 2018.

Kryvoruchko, I. S., S. Sinharoy, I. Torres-Jerez, D. Sosso, C. I. Pislariu, D. Guan, J. Murray, V. A. Benedito, W. B. Frommer and M. K. Udvardi, "MtSWEET11, a Nodule-Specific Sucrose Transporter of Medicago Truncatula", *Plant Physiology*, Vol. 171, No. 1, pp. 554–565, 2016.

Kumar, D., A. K. Yadav and D. Dash, "Choosing an Optimal Database for Protein Identification from Tandem Mass Spectrometry Data", *Methods in Molecular Biology*, Vol. 1549, No. 1, pp. 17–29, 2017.

Kute, P. M., O. Soukarieh, H. Tjeldnes, D. A. Trégouët and E. Valen, "Small Open Reading Frames, How to Find Them and Determine Their Function", *Frontiers in Genetics*, Vol. 12, No. 1, p. 2903, 2022.

Ladoukakis, E., V. Pereira, E. G. Magny, A. Eyre-Walker and J. P. Couso, "Hundreds of Putatively Functional Small Open Reading Frames in Drosophila", *Genome Biology*,

Vol. 12, No. 11, pp. 1–17, 2011.

Lasda, E. L., and T. Blumenthal, "Trans-Splicing", *Wiley Interdisciplinary Reviews: RNA*, Vol. 2, No. 3, pp. 417–434, Wiley Interdiscip Rev RNA, 2011.

Laura Howes, "Many of Our Proteins Remain Hidden in the Dark Proteome", *Chemical & Engineering News*, pp. 24–28, 2022.

Lauressergues, D., J. M. Couzigou, H. San Clemente, Y. Martinez, C. Dunand, G. Bécard and J. P. Combier, "Primary Transcripts of MicroRNAs Encode Regulatory Peptides", *Nature*, Vol. 520, No. 7545, pp. 90–93, 2015.

Leblanc, S., and M. A. Brunet, "Modelling of Pathogen-Host Systems Using Deeper ORF Annotations and Transcriptomics to Inform Proteomics Analyses", *Computational and Structural Biotechnology Journal*, Vol. 18, No. 1, pp. 2836–2850, 2020.

Lee, H. K., K. S. Mysore and J. Wen, "Tnt1 Insertional Mutagenesis in Medicago Truncatula", *Methods in Molecular Biology*, Vol. 1822, pp. 107–114, Methods in Molecular Biology, 2018.

Li, H., Y. S. Joh, H. Kim, E. Paek, S.-W. Lee and K.-B. Hwang, "Evaluating the Effect of Database Inflation in Proteogenomic Search on Sensitive and Reliable Peptide Identification", *BMC Genomics 2016 17:13*, Vol. 17, No. 13, pp. 151–162, 2016.

Li, X., H. Feng, J. Wen, J. Dong and T. Wang, "MtCAS31 Aids Symbiotic Nitrogen Fixation by Protecting the Leghemoglobin MtLb120-1 under Drought Stress in Medicago Truncatula", *Frontiers in Plant Science*, Vol. 9, No. 1, p. 633, 2018.

Limpens, E., R. Mirabella, E. Fedorova, C. Franken, H. Franssen, T. Bisseling and R. Geurts, "Formation of Organelle-like N2-Fixing Symbiosomes in Legume Root Nodules Is Controlled by DMI2", *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 102, No. 29, pp. 10375–10380, 2005.

Lin, Y., B. Tao, X. Fang, T. Wang and J. Zhang, "The Complete Mitochondrial Genome of

Lithobates Catesbeianus (Anura: Ranidae)", *Mitochondrial DNA*, Vol. 25, No. 6, pp. 447–448, 2014.

Lockwood, S., K. A. Brayton, J. A. Daily and S. L. Broschat, "Whole Proteome Clustering of 2,307 Proteobacterial Genomes Reveals Conserved Proteins and Significant Annotation Issues", *Frontiers in Microbiology*, Vol. 10, No. 1, p. 383, 2019.

Lu, C. M., Y. J. Wu, C. C. Chen, J. L. Hsu, J. C. Chen, J. Y. Chen, C. H. Huang and Y. C. Ko, "Identification of Low-Abundance Proteins via Fractionation of the Urine Proteome with Weak Anion Exchange Chromatography", *Proteome Science*, Vol. 9, No. 1, p. 17, 2011.

Ma, J., C. C. Ward, I. Jungreis, S. A. Slavoff, A. G. Schwaid, J. Neveu, B. A. Budnik, M. Kellis and A. Saghatelian, "Discovery of Human SORF-Encoded Polypeptides (SEPs) in Cell Lines and Tissue", *Journal of Proteome Research*, Vol. 13, No. 3, pp. 1757–1765, 2014.

Malhis, N., S. J. M. Jones and J. Gsponer, "Improved Measures for Evolutionary Conservation That Exploit Taxonomy Distances", *Nature Communications*, Vol. 10, No. 1, pp. 1–8, 2019.

Marx, H., C. E. Minogue, D. Jayaraman, A. L. Richards, N. W. Kwiecien, A. F. Siahpirani, S. Rajasekar, J. Maeda, K. Garcia, A. R. Del Valle-Echevarria, J. D. Volkening, M. S. Westphall, S. Roy, M. R. Sussman, J. M. Ané and J. J. Coon, "A Proteomic Atlas of the Legume Medicago Truncatula and Its Nitrogen-Fixing Endosymbiont Sinorhizobium Meliloti", *Nature Biotechnology*, Vol. 34, No. 11, pp. 1198–1205, 2016.

Marzorati, F., C. Wang, G. Pavesi, L. Mizzi and P. Morandini, "Cleaning the Medicago Microarray Database to Improve Gene Function Analysis", *Plants*, Vol. 10, No. 6, p. 1240, 2021.

McEwen, C. N., and B. S. Larsen, "Ionization Mechanisms Related to Negative Ion APPI, APCI, and DART", *Journal of the American Society for Mass Spectrometry*, Vol. 20,

No. 8, pp. 1518–1521, 2009.

Medzihradszky, K. F., and R. J. Chalkley, “Lessons in de Novo Peptide Sequencing by Tandem Mass Spectrometry”, *Mass Spectrometry Reviews*, Vol. 34, No. 1, pp. 43–63, 2015.

Menssen, A., S. Hohmann, W. Martin, P. S. Schnable, P. A. Peterson, H. Saedler and A. Gierl, “The En/Spm Transposable Element of Zea Mays Contains Splice Sites at the Termini Generating a Novel Intron from a DSpm Element in the A2 Gene”, *EMBO Journal*, Vol. 9, No. 10, pp. 3051–3057, 1990.

Mergaert, P., A. Kereszt and E. Kondorosi, “Gene Expression in Nitrogen-Fixing Symbiotic Nodule Cells in Medicago Truncatula and Other Nodulating Plants”, *The Plant Cell*, Vol. 32, No. 1, pp. 42–68, 2020.

Meydan, S., D. Klepacki, S. Karthikeyan, T. Margus, P. Thomas, J. E. Jones, Y. Khan, J. Briggs, J. D. Dinman, N. Vázquez-Laslop and A. S. Mankin, “Programmed Ribosomal Frameshifting Generates a Copper Transporter and a Copper Chaperone from the Same Gene”, *Molecular Cell*, Vol. 65, No. 2, pp. 207–219, 2017.

Meyer, J. G., “In Silico Proteome Cleavage Reveals Iterative Digestion Strategy for High Sequence Coverage”, *ISRN Computational Biology*, Vol. 2014, No. 1, pp. 1–7, 2014.

Mir, K., K. Neuhaus, S. Scherer, M. Bossert and S. Schober, “Predicting Statistical Properties of Open Reading Frames in Bacterial Genomes”, *PLoS ONE*, Vol. 7, No. 9, p. 45103, 2012.

Mittal, R. D., “Tandem Mass Spectroscopy in Diagnosis and Clinical Research”, *Indian Journal of Clinical Biochemistry*, Vol. 30, No. 2, pp. 121–123, 2015.

Mohapatra, B. R., O. Dinardo, W. D. Gould and D. W. Koren, “Genomics of Microbial Dissimilatory Reduction of Radionuclides: A Comprehensive Review”, *Reference Module in Earth Systems and Environmental Sciences*, p. 1, 2014.

Mouilleron, H., V. Delcourt and X. Roucou, “Death of a Dogma: Eukaryotic MRNAs Can Code for More than One Protein”, *Nucleic Acids Research*, Vol. 44, No. 1, pp. 14–23, 2016.

Muth, T., and B. Y. Renard, “Evaluating de Novo Sequencing in Proteomics: Already an Accurate Alternative to Database-Driven Peptide Identification?”, *Briefings in Bioinformatics*, Vol. 19, No. 5, pp. 954–970, 2018.

Naville, M., and S. Merabet, “In-Depth Annotation of the Drosophila Bithorax-Complex Reveals the Presence of Several Alternative ORFs That Could Encode for Motif-Rich Peptides”, *Cells*, Vol. 10, No. 11, p. 2983, 2021.

Neagu, A. N., M. Jayathirtha, E. Baxter, M. Donnelly, B. A. Petre and C. C. Darie, “Applications of Tandem Mass Spectrometry (MS/MS) in Protein Analysis for Biomedical Research”, *Molecules*, Vol. 27, No. 8, p. 2411, 2022.

Nelde, A., L. Flötotto, L. Jürgens, L. Szymik, E. Hubert, J. Bauer, C. Schliemann, T. Kessler, G. Lenz, H. G. Rammensee, J. S. Walz and K. Wethmar, “Upstream Open Reading Frames Regulate Translation of Cancer-Associated Transcripts and Encode HLA-Presented Immunogenic Tumor Antigens”, *Cellular and Molecular Life Sciences*, Vol. 79, No. 3, pp. 1–18, 2022.

Nibert, M. L., J. D. Pyle and A. E. Firth, “A +1 Ribosomal Frameshifting Motif Prevalent among Plant Amalgaviruses”, *Virology*, Vol. 498, No. 1, pp. 201–208, 2016.

Oldroyd, G. E. D., J. D. Murray, P. S. Poole and J. A. Downie, “The Rules of Engagement in the Legume-Rhizobial Symbiosis”, *Annual Review of Genetics*, Vol. 45, No. 1, pp. 119–144, 2011.

Olexiouk, V., and G. Menschaert, “Identification of Small Novel Coding Sequences, a Proteogenomics Endeavor”, *Advances in Experimental Medicine and Biology*, Vol. 926, No. 1, pp. 49–64, 2016.

Orr, M. W., Y. Mao, G. Storz and S. B. Qian, “Alternative ORFs and Small ORFs:

Shedding Light on the Dark Proteome", *Nucleic Acids Research*, Vol. 48, No. 3, pp. 1029–1042, 2021.

Patade, V. Y., L. C. Meher, A. Grover, S. M. Gupta and M. Nasim, "Omics Approaches in Biofuel Technologies: Toward Cost Effective, Eco-Friendly, and Renewable Energy", *Omics Technologies and Bio-engineering: Volume 2: Towards Improving Quality of Life*, pp. 337–351, Elsevier, 2018.

Pathan, M., M. Samuel, S. Keerthikumar and S. Mathivanan, "Unassigned MS/MS Spectra: Who Am I?", *Methods in Molecular Biology*, Vol. 1549, No. 1, pp. 67–74, 2017.

Pecrix, Y., S. E. Staton, E. Sallet, C. Lelandais-Brière, S. Moreau, S. Carrère, T. Blein, M. F. Jardinaud, D. Latrasse, M. Zouine, M. Zahm, J. Kreplak, B. Mayjonade, C. Satgé, M. Perez, S. Cauet, W. Marande, C. Chantry-Darmon, C. Lopez-Roques, O. Bouchez, A. Bérard, F. Debellé, S. Muños, A. Bendahmane, H. Bergès, A. Niebel, J. Buitink, F. Frugier, M. Benhamed, M. Crespi, J. Gouzy and P. Gamas, "Whole-Genome Landscape of Medicago Truncatula Symbiotic Genes", *Nature Plants*, Vol. 4, No. 12, pp. 1017–1025, 2018.

Perdigão, N., J. Heinrich, C. Stolte, K. S. Sabir, M. J. Buckley, B. Tabor, B. Signal, B. S. Gloss, C. J. Hammang, B. Rost, A. Schafferhans and S. I. O'donoghue, "Unexpected Features of the Dark Proteome", *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 112, No. 52, pp. 15898–15903, 2015.

Perdigão, N., and A. Rosa, "Dark Proteome Database: Studies on Dark Proteins", *High-Throughput*, Vol. 8, No. 2, p. 8, 2019.

Peselis, A., and A. Serganov, "Structure and Function of Pseudoknots Involved in Gene Expression Control", *Wiley Interdisciplinary Reviews. RNA*, Vol. 5, No. 6, pp. 803–822, 2014.

Pevzner, P. A., Z. Mulyukov, V. Dancik and C. L. Tang, "Efficiency of Database Search for Identification of Mutated and Modified Proteins via Mass Spectrometry", *Genome*

Research, Vol. 11, No. 2, pp. 290–299, 2001.

Pickett, B. E., R. Striker and E. J. Lefkowitz, "Evidence for Separation of HCV Subtype 1a into Two Distinct Clades", *Journal of Viral Hepatitis*, Vol. 18, No. 9, pp. 608–618, 2011.

Pienaar, E., and H. J. Viljoen, "The Tri-Frame Model", *Journal of Theoretical Biology*, Vol. 251, No. 4, pp. 616–627, 2008.

Qin, Q., S. Delrio, J. Wan, R. Jay Widmer, P. Cohen, L. O. Lerman and A. Lerman, "Downregulation of Circulating MOTS-c Levels in Patients with Coronary Endothelial Dysfunction", *International Journal of Cardiology*, Vol. 254, No. 1, pp. 23–27, 2018.

Raj, A., S. H. Wang, H. Shim, A. Harpak, Y. I. Li, B. Engelmann, M. Stephens, Y. Gilad and J. K. Pritchard, "Thousands of Novel Translated Open Reading Frames in Humans Inferred by Ribosome Footprint Profiling", *ELife*, Vol. 5, No. MAY2016, p. e13328, 2016.

Ren, P., L. Lu, S. Cai, J. Chen, W. Lin and F. Han, "Alternative Splicing: A New Cause and Potential Therapeutic Target in Autoimmune Disease", *Frontiers in Immunology*, Vol. 12, No. 1, p. 713540, 2021.

Renz, P. F., F. Valdivia Francia and A. Sendoel, "Some like It Translated: Small ORFs in the 5′UTR", *Experimental Cell Research*, Vol. 396, No. 1, p. 112229, 2020.

Ribas, J. C., and R. B. Wickner, "The Gag Domain of the Gag-Pol Fusion Protein Directs Incorporation into the L-A Double-Stranded RNA Viral Particles in Saccharomyces Cerevisiae", *Journal of Biological Chemistry*, Vol. 273, No. 15, pp. 9306–9311, 1998.

Root-Bernstein, R., and M. Root-Bernstein, "The Ribosome as a Missing Link in Prebiotic Evolution II: Ribosomes Encode Ribosomal Proteins That Bind to Common Regions of Their Own MRNAs and RRNAs", *Journal of Theoretical Biology*, Vol. 397, No. 1, pp. 115–127, 2016.

Roy, S., W. Liu, R. S. Nandety, A. Crook, K. S. Mysore, C. I. Pislariu, J. Frugoli, R. Dickstein and M. K. Udvardi, "Celebrating 20 Years of Genetic Discoveries in Legume Nodulation and Symbiotic Nitrogen Fixation", *The Plant Cell*, Vol. 32, No. 1, pp. 15–41, 2020.

Sainju, U., R. Ghimire and G. P. Pradhan, "Nitrogen Fertilization I: Impact on Crop, Soil, and Environment", *Nitrogen Fixation*, IntechOpen, 2020.

Saito, H., "The RNA World 'Hypothesis'", *Nature Reviews Molecular Cell Biology*, Vol. 23, No. 9, p. 582, 2022.

Samandi, S., A. V. Roy, V. Delcourt, J. F. Lucier, J. Gagnon, M. C. Beaudoin, B. Vanderperre, M. A. Breton, J. Motard, J. F. Jacques, M. Brunelle, I. Gagnon-Arsenault, I. Fournier, A. Ouangraoua, D. J. Hunting, A. A. Cohen, C. R. Landry, M. S. Scott and X. Roucou, "Deep Transcriptome Annotation Enables the Discovery and Functional Characterization of Cryptic Small Proteins", *ELife*, Vol. 6, No. 1, p. e27860, 2017.

Santi, C., D. Bogusz and C. Franche, "Biological Nitrogen Fixation in Non-Legume Plants", *Annals of Botany*, Vol. 111, No. 5, pp. 743–767, 2013.

Santos, M. D. M., D. B. Lima, J. S. G. Fischer, M. A. Clasen, L. U. Kurt, A. C. Camillo-Andrade, L. C. Monteiro, P. F. de Aquino, A. G. C. Neves-Ferreira, R. H. Valente, M. R. O. Trugilho, G. V. F. Brunoro, T. A. C. B. Souza, R. M. Santos, M. Batista, F. C. Gozzo, R. Durán, J. R. Yates, V. C. Barbosa and P. C. Carvalho, "Simple, Efficient and Thorough Shotgun Proteomic Analysis with PatternLab V", *Nature Protocols*, Vol. 17, No. 7, pp. 1553–1578, 2022.

Savaryn, J. P., T. K. Toby and N. L. Kelleher, "A Researcher's Guide to Mass Spectrometry-Based Proteomics", *Proteomics*, Vol. 16, No. 18, pp. 2435–2443, 2016.

Schlötterer, C., "Genes from Scratch - the Evolutionary Fate of de Novo Genes", *Trends in Genetics*, Vol. 31, No. 4, pp. 215–219, Elsevier, 2015.

Schon, K. R., R. Horvath, W. Wei, C. Calabrese, A. Tucci, K. Ibañez, T. Ratnaike, R. D. S. Pitceathly, E. Bugiardini, R. Quinlivan, M. G. Hanna, E. Clement, E. Ashton, J. A. Sayer, P. Brennan, D. Josifova, L. Izatt, C. Fratter, V. Nesbitt, T. Barrett, D. J. McMullen, A. Smith, C. Deshpande, S. F. Smithson, R. Festenstein, N. Canham, M. Caulfield, H. Houlden, S. Rahman, P. F. Chinnery, J. C. Ambrose, P. Arumugam, R. Bevers, M. Bleda, F. Boardman-Pretty, C. R. Boustred, H. Brittain, M. J. Caulfield, G. C. Chan, G. Elgar, T. Fowler, A. Giess, A. Hamblin, S. Henderson, T. J. P. Hubbard, R. Jackson, L. J. Jones, D. Kasperaviciute, M. Kayikci, A. Kousathanas, L. Lahnstein, S. E. A. Leigh, I. U. S. Leong, F. J. Lopez, F. Maleady-Crowe, M. McEntegart, F. Minneci, L. Moutsianas, M. Mueller, N. Murugaesu, A. C. Need, P. O'Donovan, C. A. Odhams, C. Patch, M. B. Pereira, D. Perez-Gil, J. Pullinger, T. Rahim, A. Rendon, T. Rogers, K. Savage, K. Sawant, R. H. Scott, A. Siddiq, A. Sieghart, S. C. Smith, A. Sosinsky, A. Stuckey, M. Tanguy, A. L. Taylor Tavares, E. R. A. Thomas, S. R. Thompson, A. Tucci, M. J. Welland, E. Williams, K. A. Witkowska and S. M. Wood, "Use of Whole Genome Sequencing to Determine Genetic Basis of Suspected Mitochondrial Disorders: Cohort Study", *The BMJ*, Vol. 375, No. 1, p. e066288, 2021.

Schwarzenbach, H., "Loss of Heterozygosity Brenner's Encyclopedia of Genetics, 271–273.", S. Maloy & K. Hughes Eds., *Brenner's Encyclopedia of Genetics (Second Edition)*, pp. 271–273, Academic Press, 2013.

Sharma, A., P. Kamal Badola, C. Bhatia, D. Sharma, # & Prabodh and K. Trivedi, "MiRNA-Encoded Peptide, MiPEP858, Regulates Plant Growth and Development in Arabidopsis", *BioRxiv*, p. 642561, 2019.

Shin, J., H. Marx, A. Richards, D. Vaneechoutte, D. Jayaraman, J. Maeda, S. Chakraborty, M. Sussman, K. Vandepoele, J. M. Ané, J. Coon and S. Roy, "A Network-Based Comparative Framework to Study Conservation and Divergence of Proteomes in Plant Phylogenies", *Nucleic Acids Research*, Vol. 49, No. 1, p. e3, 2021.

Shteynberg, D., A. I. Nesvizhskii, R. L. Moritz and E. W. Deutsch, "Combining Results of Multiple Search Engines in Proteomics", *Molecular and Cellular Proteomics*, Vol.

12, No. 9, pp. 2383–2393, 2013.

Singh, P., and E. P. Ahi, “The Importance of Alternative Splicing in Adaptive Evolution”, *Molecular Ecology*, Vol. 31, No. 7, pp. 1928–1938, John Wiley & Sons, Ltd, 2022.

Sinharoy, S., I. Torres-Jerez, K. Bandyopadhyay, A. Kereszt, C. I. Pislariu, J. Nakashima, V. A. Benedito, E. Kondorosi and M. K. Udvardi, “The C2H2 Transcription Factor REGULATOR OF SYMBIOSOME DIFFERENTIATION Represses Transcription of the Secretory Pathway Gene VAMP721a and Promotes Symbiosome Development in Medicago Truncatula”, *Plant Cell*, Vol. 25, No. 9, pp. 3584–3601, 2013.

Siuzdak, G., “An Introduction to Mass Spectrometry Ionization: An Excerpt from The Expanding Role of Mass Spectrometry in Biotechnology, 2nd Ed.; MCC Press: San Diego, 2005”, *JALA: Journal of the Association for Laboratory Automation*, Vol. 9, No. 2, pp. 50–63, 2004.

Sleno, L., and D. A. Volmer, “Ion Activation Methods for Tandem Mass Spectrometry”, *Journal of Mass Spectrometry*, Vol. 39, No. 10, pp. 1091–1112, 2004.

Smith, J. S., and R. A. Thakur, “Mass Spectrometry”, *Food Analysis*, pp. 457–470, Springer US, 2010.

Starker, C. G., A. L. Parra-Colmenares, L. Smith, R. M. Mitra and S. R. Long, “Nitrogen Fixation Mutants of Medicago Truncatula Fail to Support Plant and Bacterial Symbiotic Gene Expression”, *Plant Physiology*, Vol. 140, No. 2, pp. 671–680, 2006.

Steward, C. A., A. P. J. Parker, B. A. Minassian, S. M. Sisodiya, A. Frankish and J. Harrow, “Genome Annotation for Clinical Genomic Diagnostics: Strengths and Weaknesses”, *Genome Medicine*, Vol. 9, No. 1, pp. 1–19, 2017.

Sticker, A., L. Martens and L. Clement, “Mass Spectrometrists Should Search for All Peptides, but Assess Only the Ones They Care About”, *Nature Methods*, Vol. 14, No. 7, pp. 643–644, 2017.

Sun, Y. M., and Y. Q. Chen, “Principles and Innovative Technologies for Decrypting Noncoding RNAs: From Discovery and Functional Prediction to Clinical Application”, *Journal of Hematology and Oncology*, Vol. 13, No. 1, pp. 1–27, 2020.

Swaney, D. L., C. D. Wenger and J. J. Coon, “Value of Using Multiple Proteases for Large-Scale Mass Spectrometry-Based Proteomics”, *Journal of Proteome Research*, Vol. 9, No. 3, pp. 1323–1329, 2010.

Tariq, M. U., M. Haseeb, M. Aledhari, R. Razzak, R. M. Parizi and F. Saeed, “Methods for Proteogenomics Data Analysis, Challenges, and Scalability Bottlenecks: A Survey”, *IEEE Access*, Vol. 9, No. 1, pp. 5497–5516, 2021.

Tejada-Jiménez, M., R. Castro-Rodríguez, I. Kryvoruchko, M. Mercedes Lucas, M. Udvardi, J. Imperial and M. González-Guerrero, “Medicago Truncatula Natural Resistance-Associated Macrophage Protein1 Is Required for Iron Uptake by Rhizobia-Infected Nodule Cells”, *Plant Physiology*, Vol. 168, No. 1, pp. 258–272, 2015.

Travella, S., and B. Keller, “Down-Regulation of Gene Expression by RNA-Induced Gene Silencing”, *Methods in Molecular Biology*, Vol. 478, No. 1, pp. 185–199, 2009.

Van De Velde, W., G. Zehirov, A. Szatmari, M. Debreczeny, H. Ishihara, Z. Kevei, A. Farkas, K. Mikulass, A. Nagy, H. Tiricz, B. Satiat-Jeunemaître, B. Alunni, M. Bourge, K. I. Kucho, M. Abe, A. Kereszt, G. Maroti, T. Uchiumi, E. Kondorosi and P. Mergaert, “Plant Peptides Govern Terminal Differentiation of Bacteria in Symbiosis”, *Science*, Vol. 327, No. 5969, pp. 1122–1126, 2010.

Van Der Horst, S., T. Filipovska, J. Hanson and S. Smeekens, “Metabolite Control of Translation by Conserved Peptide UORFs: The Ribosome as a Metabolite Multisensor”, *Plant Physiology*, Vol. 182, No. 1, pp. 110–122, 2020.

VanBogelen, R. A., and F. C. Neidhardt, “Ribosomes as Sensors of Heat and Cold Shock in Escherichia Coli”, *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 87, No. 15, pp. 5589–5593, 1990.

Vanderperre, B., J. F. Lucier, C. Bissonnette, J. Motard, G. Tremblay, S. Vanderperre, M. Wisztorski, M. Salzet, F. M. Boisvert and X. Roucou, "Direct Detection of Alternative Open Reading Frames Translation Products in Human Significantly Expands the Proteome", *PLoS ONE*, Vol. 8, No. 8, p. e70698, 2013.

Varagona, M. J., M. Purugganan and S. R. Wessler, "Alternative Splicing Induced by Insertion of Retrotransposons into the Maize Waxy Gene", *Plant Cell*, Vol. 4, No. 7, pp. 811–820, 1992.

Vaudel, M., J. M. Burkhart, R. P. Zahedi, E. Oveland, F. S. Berven, A. Sickmann, L. Martens and H. Barsnes, "PeptideShaker Enables Reanalysis of MS-Derived Proteomics Data Sets: To the Editor", *Nature Biotechnology*, Vol. 33, No. 1, pp. 22–24, 2015.

Vazquez-Anderson, J., and L. M. Contreras, "Regulatory RNAs: Charming Gene Management Styles for Synthetic Biology Applications", *RNA Biology*, Vol. 10, No. 12, p. 1778, 2013.

Vekey, K., A. Telekes and A. Vertes, *Medical Applications of Mass Spectrometry*, Elsevier Science, 2011.

Vigneron, N., V. Ferrari, V. Stroobant, J. A. Habib and B. J. Van Den Eynde, "Peptide Splicing by the Proteasome", *Journal of Biological Chemistry*, Vol. 292, No. 51, pp. 21170–21179, 2017.

Von Arnim, A. G., Q. Jia and J. N. Vaughn, "Regulation of Plant Translation by Upstream Open Reading Frames", *Plant Science*, Vol. 214, No. 1, pp. 1–12, 2014.

Wan, X., J. Hontelez, A. Lillo, C. Guarnerio, D. Van De Peut, E. Fedorova, T. Bisseling and H. Franssen, "Medicago Truncatula ENOD40-1 and ENOD40-2 Are Both Involved in Nodule Initiation and Bacteroid Development", *Journal of Experimental Botany*, Vol. 58, No. 8, pp. 2033–2041, 2007.

Wang, B., V. Kumar, A. Olson and D. Ware, "Reviving the Transcriptome Studies: An

Insight into the Emergence of Single-Molecule Transcriptome Sequencing", *Frontiers in Genetics*, Vol. 10, No. APR, p. 384, 2019.

Wang, P., and S. R. Wilson, "Mass Spectrometry-Based Protein Identification by Integrating de Novo Sequencing with Database Searching.", *BMC Bioinformatics*, Vol. 14 Suppl 2, No. 2, pp. 1–9, 2013.

Watson, B. S., V. S. Asirvatham, L. Wang and L. W. Sumner, "Mapping the Proteome of Barrel Medic (Medicago Truncatula)", *Plant Physiology*, Vol. 131, No. 3, pp. 1104–1123, 2003.

Wecker, L., and J. Krzanowski, "Drug Development", *xPharm: The Comprehensive Pharmacology Reference*, pp. 1–3, Elsevier, 2007.

Weiss, R. B., D. M. Dunn, J. F. Atkins and R. F. Gesteland, "Slippery Runs, Shifty Stops, Backward Steps, and Forward Hops: -2, -1, +1, +2, +5, and +6 Ribosomal Frameshifting", *Cold Spring Harbor Symposia on Quantitative Biology*, Vol. 52, No. 1, pp. 687–693, 1987.

Williams, M., A. I. Louw and L. M. Birkholtz, "Deletion Mutagenesis of Large Areas in Plasmodium Falciparum Genes: A Comparative Study", *Malaria Journal*, Vol. 6, No. 1, pp. 1–9, 2007.

Xiao, Y., M. M. Vecchi and D. Wen, "Distinguishing between Leucine and Isoleucine by Integrated LC-MS Analysis Using an Orbitrap Fusion Mass Spectrometer", *Analytical Chemistry*, Vol. 88, No. 21, pp. 10757–10766, 2016.

Xu, C., and B. Ma, "Software for Computational Peptide Identification from MS-MS Data", *Drug Discovery Today*, Vol. 11, No. 13–14, pp. 595–600, 2006.

Xu, H., P. Wang, Y. Fu, Y. Zheng, Q. Tang, L. Si, J. You, Z. Zhang, Y. Zhu, L. Zhou, Z. Wei, B. Lin, L. Hu and X. Kong, "Length of the ORF, Position of the First AUG and the Kozak Motif Are Important Factors in Potential Dual-Coding Transcripts", *Cell Research*, Vol. 20, No. 4, pp. 445–457, 2010.

Xu, P., X. Dai, D. Wang, Y. Miao, X. Zhang, S. Wang, L. Teng, B. Dong, Z. Bao, S. Wang, Q. Lyu and W. Liu, "The Discovered Chimeric Protein Plays the Cohesive Role to Maintain Scallop Byssal Root Structural Integrity", *Scientific Reports*, Vol. 8, No. 1, pp. 1–9, 2018.

Xu, Z., J. Choi, T. S. B. Yen, W. Lu, A. Strohecker, S. Govindarajan, D. Chien, M. J. Selby and J. H. Ou, "Synthesis of a Novel Hepatitis C Virus Protein by Ribosomal Frameshift", *EMBO Journal*, Vol. 20, No. 14, pp. 3840–3848, 2001.

Xue, S., and M. Barna, "Specialized Ribosomes: A New Frontier in Gene Regulation and Organismal Biology", *Nature Reviews Molecular Cell Biology*, Vol. 13, No. 6, pp. 355–369, 2012.

Yan, S., J. Der Wen, C. Bustamante and I. Tinoco, "Ribosome Excursions during MRNA Translocation Mediate Broad Branching of Frameshift Pathways", *Cell*, Vol. 160, No. 5, pp. 870–881, 2015.

Yang, H., Y. C. Li, M. Z. Zhao, F. L. Wu, X. Wang, W. Di Xiao, Y. H. Wang, J. L. Zhang, F. Q. Wang, F. Xu, W. F. Zeng, C. M. Overall, S. M. He, H. Chi and P. Xu, "Precision de Novo Peptide Sequencing Using Mirror Proteases of Ac-Lysarginase and Trypsin for Large-Scale Proteomics", *Molecular and Cellular Proteomics*, Vol. 18, No. 4, pp. 773–785, 2019.

Yazhini, A., "Small Open Reading Frames: Tiny Treasures of the Non-Coding Genomic Regions", *Resonance*, Vol. 23, No. 1, pp. 57–67, 2018.

Yoshikawa, M., T. Iki, H. Numa, K. Miyashita, T. Meshi and M. Ishikawa, "A Short Open Reading Frame Encompassing the MicroRNA173 Target Site Plays a Role in Trans-Acting Small Interfering RNA Biogenesis", *Plant Physiology*, Vol. 171, No. 1, pp. 359–368, 2016.

Yu, L., Q. Di, D. Zhang, Y. Liu, X. Li, K. S. Mysore, J. Wen, J. Yan and L. Luo, "A Legume-Specific Novel Type of Phytosulfokine, PSK-δ, Promotes Nodulation by Enhancing Nodule Organogenesis", *Journal of Experimental Botany*, Vol. 73, No. 8,

pp. 2698–2713, 2022.

Yu, L., W. Zhou, D. Zhang, J. Yan and L. Luo, "Phytosulfokine-α Promotes Root Growth by Repressing Expression of Pectin Methylesterase Inhibitor (PMEI) Genes in Medicago Truncatula", *Phyton*, Vol. 89, No. 4, pp. 873–881, 2020.

Zhang, J., L. Yue, X. Wu, H. Liu and W. Wang, "Function of Small Peptides During Male-Female Crosstalk in Plants", *Frontiers in Plant Science*, Vol. 12, No. 1, p. 738, 2021.

Zhang, M., N. Huang, X. Yang, J. Luo, S. Yan, F. Xiao, W. Chen, X. Gao, K. Zhao, H. Zhou, Z. Li, L. Ming, B. Xie and N. Zhang, "A Novel Protein Encoded by the Circular Form of the SHPRH Gene Suppresses Glioma Tumorigenesis", *Oncogene*, Vol. 37, No. 13, pp. 1805–1814, 2018.

Zhang, W., and X. Zhao, "Method for Rapid Protein Identification in a Large Database", *BioMed Research International*, Vol. 2013, No. 1, p. 414069, 2013.

Zhang, Z., M. Burke, Y. A. Mirokhin, D. V. Tchekhovskoi, S. P. Markey, W. Yu, R. Chaerkady, S. Hess and S. E. Stein, "Reverse and Random Decoy Methods for False Discovery Rate Estimation in High Mass Accuracy Peptide Spectral Library Searches", *Journal of Proteome Research*, Vol. 17, No. 2, pp. 846–857, 2018.

Zheng, X., L. Chen, Y. Zhou, Q. Wang, Z. Zheng, B. Xu, C. Wu, Q. Zhou, W. Hu, C. Wu and J. Jiang, "A Novel Protein Encoded by a Circular RNA CircPPP1R12A Promotes Tumor Pathogenesis and Metastasis of Colon Cancer via Hippo-YAP Signaling", *Molecular Cancer*, Vol. 18, No. 1, pp. 1–13, 2019.

Zhou, B., H. Yang, C. Yang, Y. lu Bao, S. ming Yang, J. Liu and Y. feng Xiao, "Translation of Noncoding RNAs and Cancer", *Cancer Letters*, Vol. 497, No. 1, pp. 89–99, 2021.

Zhou, C., L. Han, Y. Zhao, H. Wang, J. Nakashima, J. Tong, L. Xiao and Z. Y. Wang, "Transforming Compound Leaf Patterning by Manipulating REVOLUTA in Medicago Truncatula", *Plant Journal*, Vol. 100, No. 3, pp. 562–571, 2019.

Zhu, Y., L. M. Orre, H. J. Johansson, M. Huss, J. Boekel, M. Vesterlund, A. Fernandez-Woodbridge, R. M. M. Branca and J. Lehtiö, “Discovery of Coding Regions in the Human Genome by Integrated Proteogenomics Analysis Workflow”, *Nature Communications*, Vol. 9, No. 1, pp. 1–14, 2018.

Zlotorynski, E., “The Functions of Short ORFs and Their Microproteins”, *Nature Reviews Molecular Cell Biology*, Vol. 21, No. 5, pp. 252–253, 2020.

APPENDIX A: COMPARISON OF % IDENTITY VALUES OF ALTPROTS IN THE PROTEIN SIMILARITY SEARCH AMONG DIFFERENT RNA GROUPS

Table A.1. Descriptive statistics on % identity values of altProts in the protein similarity search (BLASTP).

	n	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
mRNA	1,3427	77.032	16.7257	0.1443	76.749	77.315	24	100
ncRNA	4,398	83.785	14.9247	0.2251	83.344	84.226	25.5	100
rRNA	392	91.632	7.9183	0.3999	90.846	92.419	56.1	100
tRNA	426	97.234	3.5708	0.173	96.894	97.574	80.8	100
Total	18,643	79.394	16.5713	0.1214	79.156	79.632	24	100

Table A.2. Test for homogeneity of variances of % identity values of altProts in the protein similarity search (BLASTP).

	Levene Statistic	df1	df2	p-value
Based on Mean	386.533	3	18639	0.00
Based on Median	373.51	3	18639	0.00
Based on Median and with adjusted df	373.51	3	17735.17	0.00
Based on trimmed mean	388.739	3	18639	0.00

Table A.3. ANOVA test for % identity values of altProts in the protein similarity search (BLASTP).

	Sum of Squares	df	Mean Square	F	p-value
Between Groups	353,970.587	3	117,990.2	461.511	0.00
Within Groups	4,765,254.418	18,639	255.66		
Total	5,119,225.005	18,642			

Table A.4. Tamhane multiple comparisons test of % identity values of altProts in the protein similarity search (BLASTP).

(I) RNA Type	(J) RNA Type	Mean Difference (I-J)	Std. Error	p-value	95% Confidence Interval	
					Lower Bound	Upper Bound
mRNA	ncRNA	-6.7526*	0.2674	0.00	-7.456	-6.049
	rRNA	-14.6001*	0.4252	0.00	-15.723	-13.477
	tRNA	-20.2012*	0.2253	0.00	-20.795	-19.607
ncRNA	mRNA	6.7526*	0.2674	0.00	6.049	7.456
	rRNA	-7.8475*	0.4589	0.00	-9.058	-6.637
	tRNA	-13.4487*	0.2839	0.00	-14.196	-12.701
rRNA	mRNA	14.6001*	0.4252	0.00	13.477	15.723
	ncRNA	7.8475*	0.4589	0.00	6.637	9.058
	tRNA	-5.6012*	0.4358	0.00	-6.752	-4.45
tRNA	mRNA	20.2012*	0.2253	0.00	19.607	20.795
	ncRNA	13.4487*	0.2839	0.00	12.701	14.196
	rRNA	5.6012*	0.4358	0.00	4.45	6.752
*The mean difference is significant at the 0.05 level.						

APPENDIX B: LIST OF CANDIDATE ALTPROTS WITH TOP HIT % IDENTITY 70% OR ABOVE

The following four tables show altProts with the top hit % identity at least 70% or above. Those candidate altProts are sorted in descending order of the number of hits, and only the top 100 altProts are shown in the following tables. For the whole list of candidate altProts, see section 2.7 Data Availability.

Table B.1. mRNA-derived altProts with at least one hit in the global BLASTP analysis (e-value ≤ 0.001; % identity ≥ 70). The entries are sorted by the number of hits, and only the top 100 altProts are shown.

#	AltProt Identifier	Number of hits
1	MtrunA17_Chr7g0276191_1F_2128-2388_261	153,028
2	MtrunA17_Chr4g0047501_3F_471-791_321	140,437
3	MtrunA17_Chr6g0477601_1F_205-549_345	119,995
4	MtrunA17_Chr6g0453561_3F_78-317_240	115,226
5	MtrunA17_Chr7g0219861_1F_412-894_483	60,779
6	MtrunA17_Chr4g0053981_2F_1097-1342_246	50,886
7	MtrunA17_Chr3g0110601_3F_3594-4916_1323	50,758
8	MtrunA17_Chr7g0258421_3F_1119-1319_201	42,922
9	MtrunA17_Chr8g0335101_3F_3-329_327	33,671
10	MtrunA17_Chr8g0372571_3F_402-569_168	31,323
11	MtrunA17_Chr6g0484101_3F_39-215_177	29,350
12	MtrunA17_Chr5g0442501_3F_939-1640_702	27,561
13	MtrunA17_Chr7g0226861_1F_2716-2979_264	25,768
14	MtrunA17_Chr6g0468911_2F_710-1057_348	23,010
15	MtrunA17_Chr2g0299891_1F_253-543_291	22,799
16	MtrunA17_Chr3g0084281_3F_3369-3725_357	21,121
17	MtrunA17_Chr3g0145571_2F_497-664_168	18,703
18	MtrunA17_Chr8g0377281_1F_1804-2376_573	17,363
19	MtrunA17_Chr3g0083081_1F_2707-3297_591	16,975
20	MtrunA17_Chr4g0009321_2F_2-334_333	16,939
21	MtrunA17_Chr6g0483441_3F_1410-1826_417	16,849
22	MtrunA17_Chr2g0319221_2F_95-379_285	16,325

Table B.1. mRNA-derived altProts with at least one hit in the global BLASTP analysis (e-value ≤ 0.001; % identity ≥ 70). The entries are sorted by the number of hits, and only the top 100 altProts are shown. (cont.)

#	AltProt Identifier	Number of hits
23	MtrunA17_Chr2g0313291_2F_761-1045_285	16,240
24	MtrunA17_Chr8g0373061_3F_1170-1571_402	16,213
25	MtrunA17_Chr8g0354301_3F_129-395_267	16,060
26	MtrunA17_Chr7g0244451_2F_2096-2395_300	15,429
27	MtrunA17_Chr1g0201931_1F_1087-1443_357	15,353
28	MtrunA17_Chr7g0254581_1F_1-366_366	14,797
29	MtrunA17_Chr4g0009054_1F_3370-3927_558	14,113
30	MtrunA17_Chr3g0136001_3F_489-800_312	13,886
31	MtrunA17_Chr4g0071691_3F_924-1253_330	13,867
32	MtrunA17_Chr2g0304391_1F_31-372_342	13,676
33	MtrunA17_Chr5g0417451_3F_36-428_393	13,374
34	MtrunA17_Chr3g0107231_2F_1973-2254_282	12,799
35	MtrunA17_Chr2g0279301_1F_2185-2799_615	12,790
36	MtrunA17_Chr3g0102251_2F_1142-1684_543	12,501
37	MtrunA17_Chr5g0430111_3F_1275-1439_165	12,384
38	MtrunA17_Chr3g0083781_2F_3728-3940_213	12,301
39	MtrunA17_Chr1g0201931_2F_308-613_306	12,021
40	MtrunA17_Chr1g0162991_2F_2-442_441	11,848
41	MtrunA17_Chr4g0020391_1F_754-1113_360	11,658
42	MtrunA17_Chr6g0479781_1F_688-1131_444	11,610
43	MtrunA17_Chr2g0279301_2F_3179-3541_363	11,256
44	MtrunA17_Chr3g0107571_3F_3-269_267	11,161
45	MtrunA17_Chr3g0100781_3F_1347-1589_243	11,154
46	MtrunA17_Chr6g0477601_1F_1-135_135	10,856
47	MtrunA17_Chr6g0480971_3F_1359-2069_711	10,811
48	MtrunA17_Chr5g0416531_3F_2103-2609_507	10,573
49	MtrunA17_Chr2g0279301_1F_2893-3261_369	10,524
50	MtrunA17_Chr4g0052121_1F_982-1161_180	10,339
51	MtrunA17_Chr6g0455161_2F_959-1402_444	10,328
52	MtrunA17_Chr3g0144591_1F_1600-2079_480	10,168
53	MtrunA17_Chr7g0236411_2F_245-691_447	10,113
54	MtrunA17_CPg0492421_3F_3-254_252	9,904
55	MtrunA17_Chr6g0480411_3F_3-653_651	9,792
56	MtrunA17_Chr1g0202521_3F_1833-2033_201	9,687

Table B.1. mRNA-derived altProts with at least one hit in the global BLASTP analysis (e-value ≤ 0.001; % identity ≥ 70). The entries are sorted by the number of hits, and only the top 100 altProts are shown. (cont.)

#	AltProt Identifier	Number of hits
57	MtrunA17_Chr1g0168011_2F_2-319_318	9,614
58	MtrunA17_Chr6g0469151_2F_1286-1798_513	9,599
59	MtrunA17_Chr4g0071261_3F_1056-1286_231	9,574
60	MtrunA17_Chr1g0202521_2F_2060-2458_399	9,539
61	MtrunA17_Chr3g0102151_3F_123-713_591	9,515
62	MtrunA17_Chr6g0480531_3F_3-575_573	9,436
63	MtrunA17_Chr6g0464081_3F_282-695_414	9,246
64	MtrunA17_Chr1g0202521_1F_2320-2727_408	9,068
65	MtrunA17_Chr6g0468191_1F_829-1149_321	9,044
66	MtrunA17_Chr8g0367821_3F_1377-1805_429	9,038
67	MtrunA17_Chr6g0478321_1F_115-336_222	8,739
68	MtrunA17_Chr4g0021691_1F_1-405_405	8,669
69	MtrunA17_Chr3g0135591_1F_454-948_495	8,650
70	MtrunA17_Chr8g0362971_2F_233-493_261	8,589
71	MtrunA17_Chr3g0095361_2F_311-673_363	8,391
72	MtrunA17_Chr3g0127971_2F_1358-1846_489	8,338
73	MtrunA17_Chr6g0480981_3F_51-443_393	8,327
74	MtrunA17_Chr6g0479751_2F_41-439_399	8,294
75	MtrunA17_Chr6g0460231_2F_530-1030_501	8,271
76	MtrunA17_Chr3g0084781_3F_147-707_561	8,250
77	MtrunA17_Chr8g0350071_1F_1-432_432	8,229
78	MtrunA17_Chr4g0005381_3F_783-1046_264	8,211
79	MtrunA17_Chr3g0144591_1F_2083-2544_462	8,184
80	MtrunA17_Chr6g0474351_3F_84-284_201	8,115
81	MtrunA17_Chr6g0482191_2F_2-442_441	8,047
82	MtrunA17_Chr6g0478321_2F_1109-1357_249	7,998
83	MtrunA17_Chr7g0244451_2F_2480-2806_327	7,994
84	MtrunA17_Chr3g0106581_1F_2065-2319_255	7,915
85	MtrunA17_Chr3g0142341_3F_39-560_522	7,871
86	MtrunA17_Chr4g0005601_3F_870-1328_459	7,838
87	MtrunA17_Chr1g0157851_3F_3-116_114	7,807
88	MtrunA17_Chr5g0408491_1F_1-204_204	7,798
89	MtrunA17_Chr7g0240791_1F_214-444_231	7,719
90	MtrunA17_Chr4g0005281_3F_495-692_198	7,578

Table B.1. mRNA-derived altProts with at least one hit in the global BLASTP analysis (e-value ≤ 0.001; % identity ≥ 70). The entries are sorted by the number of hits, and only the top 100 altProts are shown. (cont.)

#	AltProt Identifier	Number of hits
91	MtrunA17_Chr6g0482251_3F_3-338_336	7,522
92	MtrunA17_Chr5g0443951_2F_1055-1354_300	7,487
93	MtrunA17_Chr3g0084771_3F_1398-1742_345	7,410
94	MtrunA17_Chr3g0135231_2F_2876-3085_210	7,408
95	MtrunA17_Chr8g0363901_2F_200-751_552	7,397
96	MtrunA17_Chr8g0366081_3F_561-728_168	7,186
97	MtrunA17_Chr5g0425221_3F_1620-1868_249	7,126
98	MtrunA17_Chr1g0180921_3F_147-659_513	7,068
99	MtrunA17_Chr5g0443831_1F_982-1233_252	7,009
100	MtrunA17_Chr8g0342601_1F_1-420_420	6,900

Table B.2. ncRNA-derived altProts with at least one hit in the global BLASTP analysis (e-value ≤ 0.001; % identity ≥ 70). The entries are sorted by the number of hits, and only the top 100 altProts are shown.

#	AltProt Identifier	Number of hits
1	MtrunA17_Chr7g1030632_1F_28981-32112_3132	83,684
2	MtrunA17_Chr7g1030653_2F_6104-7402_1299	52,325
3	MtrunA17_Chr7g1030653_2F_5603-6100_498	36,696
4	MtrunA17_Chr3g1010907_2F_4811-5446_636	33,845
5	MtrunA17_Chr7g1030653_2F_7406-8524_1119	29,657
6	MtrunA17_Chr3g1012388_3F_1257-2423_1167	25,078
7	MtrunA17_Chr2g1008736_1F_6319-7236_918	23,688
8	MtrunA17_Chr2g1008736_1F_5113-6315_1203	21,454
9	MtrunA17_Chr7g1029142_3F_1695-2519_825	19,528
10	MtrunA17_Chr7g1029142_1F_2395-2865_471	16,664
11	MtrunA17_Chr8g1036334_3F_900-1256_357	13,554
12	MtrunA17_Chr4g0063681_2F_185-397_213	12,517
13	MtrunA17_Chr7g0228971_1F_88-282_195	11,468
14	MtrunA17_Chr3g1012388_3F_846-1253_408	10,855
15	MtrunA17_Chr4g1015884_2F_332-535_204	10,685
16	MtrunA17_Chr2g0289111_2F_290-652_363	10,487
17	MtrunA17_Chr2g0304521_2F_71-793_723	9,385

Table B.2. ncRNA-derived altProts with at least one hit in the global BLASTP analysis (e-value ≤ 0.001; % identity ≥ 70). The entries are sorted by the number of hits, and only the top 100 altProts are shown. (cont.)

#	AltProt Identifier	Number of hits
18	MtrunA17_Chr3g1011650_2F_26-340_315	9,103
19	MtrunA17_Chr2g0303491_3F_3-263_261	9,042
20	MtrunA17_Chr3g0116221_2F_2-463_462	8,971
21	MtrunA17_Chr5g0417061_2F_2-406_405	8,932
22	MtrunA17_Chr2g1008736_1F_7240-7479_240	8,705
23	MtrunA17_Chr6g0484391_2F_2-286_285	8,537
24	MtrunA17_Chr4g1018210_3F_21-332_312	8,252
25	MtrunA17_Chr5g0436421_1F_16-294_279	8,134
26	MtrunA17_Chr1g0176761_2F_2-232_231	8,091
27	MtrunA17_Chr2g1008736_2F_4808-5170_363	7,345
28	MtrunA17_Chr8g1039062_3F_849-1040_192	7,294
29	MtrunA17_Chr6g0480881_1F_1-369_369	7,276
30	MtrunA17_Chr5g1024587_2F_134-346_213	6,993
31	MtrunA17_Chr6g0464771_3F_3-548_546	6,946
32	MtrunA17_Chr7g1030632_2F_27752-29029_1278	6,547
33	MtrunA17_Chr7g1029142_1F_2869-3168_300	5,332
34	MtrunA17_Chr7g0234881_1F_280-543_264	5,208
35	MtrunA17_Chr5g0425621_2F_173-385_213	5,079
36	MtrunA17_Chr3g0094901_2F_2-256_255	4,682
37	MtrunA17_CPg0492611_2F_173-292_120	4,654
38	MtrunA17_MTg0490821_3F_204-554_351	4,450
39	MtrunA17_Chr6g1026690_2F_2003-2797_795	4,390
40	MtrunA17_Chr2g1005214_1F_697-1278_582	4,154
41	MtrunA17_Chr8g0353151_3F_3-335_333	3,951
42	MtrunA17_Chr6g0451461_2F_164-490_327	3,768
43	MtrunA17_Chr8g1036687_3F_126-2699_2574	3,756
44	MtrunA17_Chr1g0195631_1F_310-516_207	3,697
45	MtrunA17_Chr5g1023635_3F_3-650_648	3,650
46	MtrunA17_Chr7g0230941_3F_1041-1289_249	3,466
47	MtrunA17_Chr3g0116241_1F_43-228_186	3,309
48	MtrunA17_Chr8g0361321_3F_3-161_159	3,233
49	MtrunA17_Chr5g1022952_2F_332-517_186	3,137
50	MtrunA17_Chr1g1001863_3F_4002-5153_1152	3,028
51	MtrunA17_Chr1g1004507_3F_3-128_126	2,870
52	MtrunA17_Chr8g0383621_1F_61-195_135	2,622

Table B.2. ncRNA-derived altProts with at least one hit in the global BLASTP analysis (e-value ≤ 0.001; % identity ≥ 70). The entries are sorted by the number of hits, and only the top 100 altProts are shown. (cont.)

#	AltProt Identifier	Number of hits
53	MtrunA17_Chr8g0363551_1F_1-261_261	2,559
54	MtrunA17_Chr7g1029142_1F_643-885_243	2,511
55	MtrunA17_Chr2g0300091_1F_76-171_96	2,463
56	MtrunA17_Chr2g0290551_2F_2-340_339	2,387
57	MtrunA17_Chr1g1004507_3F_132-305_174	2,371
58	MtrunA17_Chr7g0256091_1F_232-372_141	2,332
59	MtrunA17_Chr4g0071171_1F_1-318_318	2,285
60	MtrunA17_Chr3g0110701_3F_144-314_171	2,127
61	MtrunA17_Chr1g0181711_2F_2-172_171	1,973
62	MtrunA17_Chr1g0164061_3F_3-602_600	1,874
63	MtrunA17_Chr4g0066871_1F_298-498_201	1,872
64	MtrunA17_Chr3g1010438_1F_2461-2883_423	1,844
65	MtrunA17_Chr4g1016073_1F_88-270_183	1,838
66	MtrunA17_Chr5g0428151_1F_43-366_324	1,824
67	MtrunA17_Chr1g1001651_3F_330-1028_699	1,722
68	MtrunA17_Chr2g0302721_3F_3-152_150	1,713
69	MtrunA17_Chr3g0093371_1F_487-597_111	1,618
70	MtrunA17_Chr7g0225081_2F_134-298_165	1,606
71	MtrunA17_Chr1g1002128_2F_2-223_222	1,577
72	MtrunA17_Chr1g1002127_1F_1858-1959_102	1,555
73	MtrunA17_Chr1g1002227_2F_4022-4957_936	1,477
74	MtrunA17_Chr2g1006500_3F_1614-1826_213	1,422
75	MtrunA17_Chr8g0343371_3F_732-896_165	1,380
76	MtrunA17_Chr2g1005214_3F_1443-1793_351	1,308
77	MtrunA17_Chr3g1010907_1F_4642-4953_312	1,297
78	MtrunA17_Chr8g0348011_3F_318-527_210	1,263
79	MtrunA17_Chr1g1002127_1F_1765-1854_90	1,261
80	MtrunA17_Chr5g1022564_3F_2202-2501_300	1,239
81	MtrunA17_Chr3g1012625_2F_2-205_204	1,213
82	MtrunA17_Chr3g0133271_1F_502-744_243	1,208
83	MtrunA17_Chr7g0250571_1F_253-420_168	1,205
84	MtrunA17_Chr7g0243161_2F_596-793_198	1,180
85	MtrunA17_Chr4g0054091_3F_3-338_336	1,150
86	MtrunA17_Chr3g0109301_3F_51-272_222	1,135
87	MtrunA17_Chr7g1031544_2F_926-1078_153	1,103

Table B.2. ncRNA-derived altProts with at least one hit in the global BLASTP analysis (e-value ≤ 0.001; % identity ≥ 70). The entries are sorted by the number of hits, and only the top 100 altProts are shown. (cont.)

#	AltProt Identifier	Number of hits
88	MtrunA17_Chr4g0012301_1F_61-348_288	1,040
89	MtrunA17_Chr2g1007624_1F_4885-5046_162	1,009
90	MtrunA17_Chr4g0012451_3F_69-404_336	1,000
91	MtrunA17_Chr6g1026690_1F_928-1260_333	976
92	MtrunA17_Chr5g1022104_1F_175-363_189	914
93	MtrunA17_Chr6g0484241_2F_383-946_564	905
94	MtrunA17_Chr8g1036687_1F_7678-7863_186	882
95	MtrunA17_Chr4g1018874_2F_107-532_426	859
96	MtrunA17_Chr7g0247671_3F_3-221_219	829
97	MtrunA17_Chr5g0417801_2F_806-1015_210	816
98	MtrunA17_Chr2g1008736_2F_4505-4804_300	811
99	MtrunA17_Chr4g1015662_2F_809-1390_582	800
100	MtrunA17_Chr2g0287661_1F_358-459_102	797

Table B.3. rRNA-derived altProts with at least one hit in the global BLASTP analysis (e-value ≤ 0.001; % identity ≥ 70). The entries are sorted by the number of hits, and only the top 100 altProts are shown.

#	AltProt Identifier	Number of hits
1	MtrunA17_Chr5g0422291_3F_5109-5639_531	660
2	MtrunA17_Chr5g0422291_2F_4808-5617_810	443
3	MtrunA17_Chr5g0422291_3F_4821-5057_237	392
4	MtrunA17_Chr5g0422291_3F_5886-6146_261	324
5	MtrunA17_Chr4g0016141_2F_5324-6058_735	244
6	MtrunA17_CPg0492331_3F_6288-6938_651	243
7	MtrunA17_Chr5g0422081_3F_2367-2540_174	205
8	MtrunA17_Chr5g0422291_2F_2666-2839_174	205
9	MtrunA17_CPg0492331_3F_1188-1502_315	202
10	MtrunA17_Chr4g0000621_2F_224-538_315	202
11	MtrunA17_Chr4g0016141_2F_224-538_315	202
12	MtrunA17_Chr7g0229401_1F_1-228_228	195

Table B.3. rRNA-derived altProts with at least one hit in the global BLASTP analysis (e-value ≤ 0.001; % identity ≥ 70). The entries are sorted by the number of hits, and only the top 100 altProts are shown. (cont.)

#	AltProt Identifier	Number of hits
13	MtrunA17_Chr5g0422271_3F_1557-1685_129	187
14	MtrunA17_Chr5g0422291_2F_6518-6658_141	178
15	MtrunA17_Chr5g0422081_2F_1940-2236_297	160
16	MtrunA17_Chr5g0422271_2F_1130-1426_297	160
17	MtrunA17_Chr5g0422291_1F_2239-2535_297	160
18	MtrunA17_Chr5g0422291_2F_5858-6016_159	160
19	MtrunA17_CPg0492331_1F_2053-2376_324	147
20	MtrunA17_Chr4g0016141_3F_1089-1412_324	147
21	MtrunA17_Chr7g0229401_2F_779-1102_324	147
22	MtrunA17_Chr5g0422081_3F_2100-2270_171	142
23	MtrunA17_Chr5g0422271_3F_1290-1460_171	142
24	MtrunA17_Chr5g0422291_2F_2399-2569_171	142
25	MtrunA17_Chr5g0422291_1F_4750-4917_168	133
26	MtrunA17_Chr5g0422041_3F_807-1037_231	130
27	MtrunA17_Chr5g0422291_1F_4546-4689_144	128
28	MtrunA17_Chr4g0000621_2F_2297-2551_255	115
29	MtrunA17_Chr5g0422041_2F_629-829_201	112
30	MtrunA17_CPg0492331_1F_3262-3471_210	110
31	MtrunA17_CPg0492331_1F_3541-3738_198	110
32	MtrunA17_Chr4g0000621_3F_2574-2771_198	110
33	MtrunA17_Chr4g0016141_3F_2298-2507_210	110
34	MtrunA17_Chr4g0016141_3F_2577-2774_198	110
35	MtrunA17_Chr7g0229401_3F_1989-2198_210	110
36	MtrunA17_Chr4g0000621_3F_1089-1328_240	107
37	MtrunA17_Chr7g0229401_3F_2268-2465_198	106
38	MtrunA17_Chr3g0090811_1F_1030-1257_228	103
39	MtrunA17_Chr4g0073121_1F_1030-1257_228	103
40	MtrunA17_CPg0492331_3F_3675-4028_354	92
41	MtrunA17_Chr4g0000621_2F_2708-3061_354	92
42	MtrunA17_Chr4g0016141_2F_2711-3064_354	92
43	MtrunA17_Chr7g0229401_2F_2402-2755_354	91
44	MtrunA17_CPg0492331_3F_2403-2627_225	88
45	MtrunA17_Chr4g0000621_1F_1438-1662_225	88
46	MtrunA17_Chr4g0016141_2F_1439-1663_225	88
47	MtrunA17_Chr7g0229401_1F_1129-1353_225	88
48	MtrunA17_Chr5g0422291_1F_6907-7119_213	86

Table B.3. rRNA-derived altProts with at least one hit in the global BLASTP analysis (e-value ≤ 0.001; % identity ≥ 70). The entries are sorted by the number of hits, and only the top 100 altProts are shown. (cont.)

#	AltProt Identifier	Number of hits
49	MtrunA17_Chr5g0422041_3F_1071-1229_159	80
50	MtrunA17_Chr5g0422271_1F_37-144_108	79
51	MtrunA17_Chr5g0422271_3F_840-1148_309	79
52	MtrunA17_Chr5g0422291_1F_37-144_108	79
53	MtrunA17_CPg0492331_3F_4449-4619_171	77
54	MtrunA17_Chr4g0000621_2F_3482-3652_171	77
55	MtrunA17_Chr4g0016141_2F_3485-3655_171	77
56	MtrunA17_Chr5g0422081_1F_886-1005_120	77
57	MtrunA17_Chr5g0422291_3F_1185-1304_120	77
58	MtrunA17_Chr7g0229401_2F_3176-3346_171	77
59	MtrunA17_Chr1g0208361_1F_1-117_117	75
60	MtrunA17_Chr3g0100621_1F_1-117_117	75
61	MtrunA17_Chr3g0090811_3F_909-1070_162	74
62	MtrunA17_Chr4g0073121_3F_909-1070_162	74
63	MtrunA17_Chr8g0359061_1F_1-117_117	74
64	MtrunA17_CPg0492331_1F_3742-3918_177	72
65	MtrunA17_Chr4g0000621_3F_2775-2951_177	72
66	MtrunA17_Chr4g0016141_3F_2778-2954_177	72
67	MtrunA17_Chr5g0422271_2F_491-742_252	72
68	MtrunA17_Chr5g0422291_1F_538-789_252	72
69	MtrunA17_Chr5g0422081_3F_1740-1958_219	71
70	MtrunA17_Chr5g0422291_2F_2039-2257_219	71
71	MtrunA17_Chr4g0000621_2F_1277-1411_135	69
72	MtrunA17_Chr7g0229401_3F_2469-2645_177	67
73	MtrunA17_CPg0492331_3F_3090-3233_144	66
74	MtrunA17_Chr4g0000621_1F_2125-2268_144	66
75	MtrunA17_Chr4g0016141_2F_2126-2269_144	66
76	MtrunA17_Chr7g0229401_2F_1817-1960_144	66
77	MtrunA17_CPg0492331_2F_2729-2929_201	65
78	MtrunA17_Chr4g0000621_3F_1764-1964_201	65
79	MtrunA17_Chr4g0016141_1F_1765-1965_201	65
80	MtrunA17_Chr5g0422041_3F_462-611_150	63
81	MtrunA17_Chr5g0422081_1F_1162-1371_210	63
82	MtrunA17_Chr5g0422291_3F_1461-1670_210	63
83	MtrunA17_Chr3g0090811_3F_1299-1454_156	62
84	MtrunA17_Chr7g0229401_3F_261-479_219	62

Table B.3. rRNA-derived altProts with at least one hit in the global BLASTP analysis (e-value ≤ 0.001; % identity ≥ 70). The entries are sorted by the number of hits, and only the top 100 altProts are shown. (cont.)

#	AltProt Identifier	Number of hits
85	MtrunA17_CPg0492331_2F_1535-1753_219	61
86	MtrunA17_Chr4g0000621_1F_571-789_219	61
87	MtrunA17_Chr4g0016141_1F_571-789_219	61
88	MtrunA17_Chr4g0073121_3F_1299-1454_156	61
89	MtrunA17_Chr5g0422081_3F_273-500_228	61
90	MtrunA17_Chr7g0229401_1F_1516-1656_141	59
91	MtrunA17_Chr5g0422041_2F_1349-1471_123	58
92	MtrunA17_Chr5g0422291_1F_979-1224_246	57
93	MtrunA17_CPg0492331_3F_3483-3668_186	56
94	MtrunA17_Chr4g0016141_2F_2519-2704_186	56
95	MtrunA17_Chr5g0422081_3F_1293-1463_171	56
96	MtrunA17_Chr5g0422291_2F_1592-1762_171	56
97	MtrunA17_Chr7g0229401_2F_2210-2395_186	56
98	MtrunA17_CPg0492331_1F_4852-4965_114	54
99	MtrunA17_Chr4g0016141_3F_3888-4001_114	54
100	MtrunA17_Chr5g0422041_3F_3-107_105	53

Table B.4. tRNA-derived altProts with at least one hit in the global BLASTP analysis (e-value ≤ 0.001; % identity ≥ 70). The entries are sorted by the number of hits, and only the top 100 altProts are shown.

#	AltProt Identifier	Number of hits
1	MtrunA17_Chr1g0156721_2F_2-85_84	65
2	MtrunA17_Chr3g0099261_2F_2-85_84	65
3	MtrunA17_Chr3g0142821_2F_5-88_84	65
4	MtrunA17_Chr4g0000301_3F_3-86_84	65
5	MtrunA17_Chr4g0000331_2F_2-85_84	65
6	MtrunA17_Chr4g0036081_2F_2-85_84	65
7	MtrunA17_Chr5g0416701_2F_2-85_84	65
8	MtrunA17_Chr5g0437871_3F_3-86_84	65
9	MtrunA17_Chr8g0344351_2F_2-85_84	65
10	MtrunA17_Chr3g0103021_3F_3-86_84	60
11	MtrunA17_Chr7g0246101_2F_2-85_84	49
12	MtrunA17_Chr5g0440611_2F_2-85_84	46
13	MtrunA17_Chr7g0231761_2F_2-85_84	44

Table B.4. tRNA-derived altProts with at least one hit in the global BLASTP analysis (e-value ≤ 0.001; % identity ≥ 70). The entries are sorted by the number of hits, and only the top 100 altProts are shown. (cont.)

#	AltProt Identifier	Number of hits
14	MtrunA17_Chr1g0208351_2F_2-76_75	35
15	MtrunA17_Chr4g0000641_2F_2-76_75	35
16	MtrunA17_Chr5g0410521_2F_2-76_75	16
17	MtrunA17_Chr3g0090911_2F_2-79_78	13
18	MtrunA17_Chr4g0016041_2F_2-79_78	13
19	MtrunA17_Chr4g0024491_2F_2-79_78	13
20	MtrunA17_Chr6g0466151_1F_1-78_78	12
21	MtrunA17_Chr1g0150931_3F_3-74_72	11
22	MtrunA17_Chr1g0198691_1F_13-87_75	11
23	MtrunA17_Chr1g0199241_2F_2-85_84	11
24	MtrunA17_Chr2g0308981_2F_2-82_81	11
25	MtrunA17_Chr2g0320701_3F_3-83_81	11
26	MtrunA17_Chr2g0329391_3F_3-83_81	11
27	MtrunA17_Chr3g0124201_1F_1-72_72	11
28	MtrunA17_Chr3g0126891_1F_1-72_72	11
29	MtrunA17_Chr3g0136191_1F_1-72_72	11
30	MtrunA17_Chr3g0141151_1F_4-75_72	11
31	MtrunA17_Chr4g0027481_2F_2-82_81	11
32	MtrunA17_Chr4g0032391_2F_2-82_81	11
33	MtrunA17_Chr4g0036381_2F_2-82_81	11
34	MtrunA17_Chr4g0036481_2F_2-82_81	11
35	MtrunA17_Chr4g0036511_2F_2-82_81	11
36	MtrunA17_Chr4g0036591_2F_2-82_81	11
37	MtrunA17_Chr4g0070401_2F_2-82_81	11
38	MtrunA17_Chr5g0405711_2F_2-73_72	11
39	MtrunA17_Chr5g0412391_3F_3-83_81	11
40	MtrunA17_Chr6g0474291_2F_2-82_81	11
41	MtrunA17_Chr7g0239781_1F_1-72_72	11
42	MtrunA17_Chr7g0258881_1F_4-84_81	11
43	MtrunA17_Chr7g0260661_1F_1-72_72	11
44	MtrunA17_Chr8g0352141_3F_3-86_84	11
45	MtrunA17_Chr8g0356231_1F_1-72_72	11
46	MtrunA17_Chr8g0366251_1F_1-72_72	11
47	MtrunA17_Chr8g0367621_2F_2-85_84	11
48	MtrunA17_Chr8g0379251_1F_1-75_75	11
49	MtrunA17_Chr1g0152271_1F_1-75_75	10
50	MtrunA17_Chr1g0172561_1F_1-75_75	10
51	MtrunA17_Chr1g0207111_2F_2-76_75	10

Table B.4. tRNA-derived altProts with at least one hit in the global BLASTP analysis (e-value ≤ 0.001; % identity ≥ 70). The entries are sorted by the number of hits, and only the top 100 altProts are shown. (cont.)

#	AltProt Identifier	Number of hits
52	MtrunA17_Chr2g0305981_1F_1-78_78	10
53	MtrunA17_Chr2g0310181_1F_1-75_75	10
54	MtrunA17_Chr2g0321021_3F_3-77_75	10
55	MtrunA17_Chr2g0321031_3F_3-77_75	10
56	MtrunA17_Chr3g0107301_1F_1-75_75	10
57	MtrunA17_Chr3g0113391_1F_1-72_72	10
58	MtrunA17_Chr5g0433581_1F_1-72_72	10
59	MtrunA17_Chr7g0245791_1F_1-72_72	10
60	MtrunA17_Chr7g0253261_1F_1-75_75	10
61	MtrunA17_Chr7g0269021_1F_1-72_72	10
62	MtrunA17_Chr8g0390791_1F_1-72_72	10
63	MtrunA17_Chr8g0390831_1F_1-72_72	10
64	MtrunA17_Chr1g0182911_3F_3-74_72	9
65	MtrunA17_Chr1g0207071_1F_1-78_78	9
66	MtrunA17_Chr2g0332201_1F_1-72_72	9
67	MtrunA17_Chr4g0036451_2F_2-82_81	9
68	MtrunA17_Chr5g0419831_1F_1-72_72	9
69	MtrunA17_Chr1g0180801_3F_3-74_72	8
70	MtrunA17_Chr1g0184321_3F_3-71_69	8
71	MtrunA17_Chr1g0194131_2F_2-79_78	8
72	MtrunA17_Chr1g0204851_2F_2-76_75	8
73	MtrunA17_Chr2g0283701_3F_3-71_69	8
74	MtrunA17_Chr2g0286141_2F_2-73_72	8
75	MtrunA17_Chr2g0288591_2F_2-73_72	8
76	MtrunA17_Chr2g0290291_2F_2-76_75	8
77	MtrunA17_Chr2g0300531_1F_4-78_75	8
78	MtrunA17_Chr2g0318531_2F_2-76_75	8
79	MtrunA17_Chr2g0324491_3F_3-71_69	8
80	MtrunA17_Chr2g0331421_2F_2-82_81	8
81	MtrunA17_Chr3g0081751_3F_3-71_69	8
82	MtrunA17_Chr3g0081801_3F_3-77_75	8
83	MtrunA17_Chr3g0105931_2F_2-82_81	8
84	MtrunA17_Chr3g0106281_3F_3-71_69	8
85	MtrunA17_Chr3g0113921_2F_2-76_75	8
86	MtrunA17_Chr3g0139451_2F_2-73_72	8
87	MtrunA17_Chr4g0023791_2F_2-82_81	8
88	MtrunA17_Chr4g0026271_3F_3-74_72	8
89	MtrunA17_Chr4g0052191_3F_3-74_72	8

Table B.4. tRNA-derived altProts with at least one hit in the global BLASTP analysis (e-value ≤ 0.001; % identity ≥ 70). The entries are sorted by the number of hits, and only the top 100 altProts are shown. (cont.)

#	AltProt Identifier	Number of hits
90	MtrunA17_Chr4g0069641_3F_3-71_69	8
91	MtrunA17_Chr4g0075061_2F_2-73_72	8
92	MtrunA17_Chr5g0404971_3F_3-74_72	8
93	MtrunA17_Chr5g0433091_3F_3-74_72	8
94	MtrunA17_Chr5g0433111_3F_3-74_72	8
95	MtrunA17_Chr5g0440941_3F_3-71_69	8
96	MtrunA17_Chr5g0444411_2F_2-79_78	8
97	MtrunA17_Chr6g0475671_3F_3-71_69	8
98	MtrunA17_Chr7g0243251_3F_3-74_72	8
99	MtrunA17_Chr7g0248761_2F_2-73_72	8
100	MtrunA17_Chr7g0271471_3F_3-80_78	8

APPENDIX C: LIST OF ALTPROTS VALIDATED BY MS SEARCHES

The following table, Table C.1, shows altProts validated by MS searches. Only altProts validated in more than one organ/condition are shown. For the whole list, please see section 2.7 Data Availability. In the following table, “m” and “n” in type column stands for mRNA and ncRNA-derived altProts.

Table C.1. AltProts validated by MS searches.

#	AltProt Identifier	Type	10-day Nodules	14-day Nodules	28-day Nodules	Buds	Flowers	Leaves	Roots	Seeds	Stems	Whole Plant	Total
1	MtrunA17_Chr3g0113931_1F_1804-1890_87	m	1	1	0	1	1	1	1	1	1	0	8
2	MtrunA17_Chr1g0157851_3F_3-116_114	m	1	1	0	1	1	1	0	1	0	1	7
3	MtrunA17_Chr1g0159811_2F_1409-1648_240	m	1	1	0	1	1	1	0	1	0	1	7
4	MtrunA17_Chr4g0038111_1F_385-480_96	m	1	1	0	1	1	1	0	0	1	1	7
5	MtrunA17_Chr4g0048901_2F_1043-1177_135	m	1	1	0	1	1	0	1	1	0	1	7
6	MtrunA17_Chr5g0410111_2F_323-568_246	m	1	1	0	1	1	0	0	1	1	1	7
7	MtrunA17_Chr8g0383131_2F_536-760_225	m	1	1	0	1	1	1	0	1	0	1	7
8	MtrunA17_Chr1g0194811_3F_582-881_300	m	1	0	0	1	1	0	0	1	0	1	5
9	MtrunA17_Chr3g0142661_1F_25-264_240	m	1	1	0	0	1	1	0	1	0	0	5
10	MtrunA17_Chr4g0052001_3F_3-302_300	m	1	0	0	1	1	0	0	0	1	1	5
11	MtrunA17_Chr5g0440901_2F_56-208_153	m	1	0	0	1	1	1	0	0	0	1	5
12	MtrunA17_Chr6g0465861_2F_467-559_93	m	0	0	0	0	1	1	1	0	1	1	5
13	MtrunA17_Chr7g0266741_3F_894-1136_243	m	1	1	0	1	1	0	0	1	0	0	5
14	MtrunA17_Chr7g0270111_3F_1794-1991_198	m	0	1	0	0	1	1	0	1	0	1	5
15	MtrunA17_CPg0492421_3F_3-254_252	m	0	0	0	1	1	1	0	1	0	1	5
16	MtrunA17_Chr1g0159791_2F_1352-1582_231	m	0	0	0	1	1	1	0	0	1	0	4
17	MtrunA17_Chr2g0310741_1F_622-681_60	m	1	1	0	1	1	0	0	0	0	0	4
18	MtrunA17_Chr3g0108701_1F_76-321_246	m	1	0	0	1	1	0	0	0	1	0	4
19	MtrunA17_Chr3g0113601_3F_708-788_81	m	0	0	0	1	1	0	0	1	0	1	4
20	MtrunA17_Chr4g0045821_3F_111-323_213	m	0	1	0	0	1	0	0	1	0	1	4
21	MtrunA17_Chr4g0072701_1F_331-447_117	n	1	0	0	0	1	0	0	1	0	1	4

Table C.1. AltProts validated by MS searches. (cont.)

#	AltProt Identifier	Type	10-day Nodules	14-day Nodules	28-day Nodules	Buds	Flowers	Leaves	Roots	Seeds	Stems	Whole Plant	Total
22	MtrunA17_Chr5g0428001_2F_2183-2278_96	m	0	0	0	1	0	1	0	0	1	1	4
23	MtrunA17_Chr6g0465901_3F_207-299_93	n	0	0	0	0	1	0	1	0	1	1	4
24	MtrunA17_Chr6g0488521_1F_1600-1674_75	m	0	1	0	0	1	1	0	0	1	0	4
25	MtrunA17_Chr7g0237621_2F_434-547_114	m	0	0	0	1	1	1	1	0	0	0	4
26	MtrunA17_Chr7g0273141_1F_625-816_192	m	1	1	0	0	0	0	1	0	0	1	4
27	MtrunA17_Chr8g0334341_3F_144-647_504	m	1	0	0	1	1	0	0	1	0	0	4
28	MtrunA17_CPg0492611_2F_173-292_120	n	0	0	0	1	1	1	0	1	0	0	4
29	MtrunA17_Chr1g0148371_1F_2233-2337_105	m	1	1	0	0	0	0	0	0	0	1	3
30	MtrunA17_Chr1g0148591_3F_1413-1622_210	m	1	1	0	0	0	0	0	0	0	1	3
31	MtrunA17_Chr1g0182591_1F_4162-4257_96	m	0	1	0	0	0	0	0	1	0	1	3
32	MtrunA17_Chr1g0185401_3F_468-650_183	m	0	0	0	1	1	0	0	1	0	0	3
33	MtrunA17_Chr1g0200501_2F_1994-2134_141	m	0	1	0	1	0	0	0	1	0	0	3
34	MtrunA17_Chr1g0204591_3F_1251-1418_168	n	1	0	0	1	0	0	0	1	0	0	3
35	MtrunA17_Chr2g0290551_2F_2-340_339	n	0	0	0	1	1	0	0	0	0	1	3
36	MtrunA17_Chr2g0294541_3F_45-353_309	m	0	0	0	1	1	0	0	0	1	0	3
37	MtrunA17_Chr2g0298731_1F_3538-3663_126	m	0	0	0	0	0	1	1	0	1	0	3
38	MtrunA17_Chr2g0303521_2F_824-949_126	m	1	1	1	0	0	0	0	0	0	0	3
39	MtrunA17_Chr2g0312631_1F_2020-2094_75	m	0	0	0	0	0	1	1	0	1	0	3
40	MtrunA17_Chr2g0324871_2F_2-271_270	m	1	0	0	0	1	0	0	0	0	1	3
41	MtrunA17_Chr3g0077951_3F_231-314_84	m	0	1	0	0	1	0	0	1	0	0	3
42	MtrunA17_Chr3g0121981_3F_240-359_120	n	0	0	0	1	1	1	0	0	0	0	3
43	MtrunA17_Chr3g0130841_2F_266-373_108	m	0	0	0	1	1	1	0	0	0	0	3
44	MtrunA17_Chr3g0132331_1F_1-75_75	m	0	0	0	1	1	1	0	0	0	0	3
45	MtrunA17_Chr3g1013840_2F_848-964_117	n	0	0	0	0	0	1	1	0	1	0	3
46	MtrunA17_Chr4g0009611_2F_413-508_96	m	0	0	0	0	1	0	0	1	0	1	3
47	MtrunA17_Chr4g0010191_2F_197-349_153	n	0	0	0	1	0	0	1	1	0	0	3
48	MtrunA17_Chr4g0021441_1F_199-303_105	m	1	1	0	0	0	0	0	1	0	0	3
49	MtrunA17_Chr4g0023881_2F_74-295_222	n	1	1	0	0	0	0	0	0	0	1	3
50	MtrunA17_Chr4g0035051_1F_652-735_84	m	0	0	0	1	0	0	0	0	1	1	3
51	MtrunA17_Chr4g0045961_3F_180-263_84	m	0	0	0	1	0	1	0	0	0	1	3
52	MtrunA17_Chr4g0046711_2F_683-1105_423	m	0	1	0	0	1	0	0	1	0	0	3
53	MtrunA17_Chr4g0050761_3F_1293-1379_87	m	0	1	0	0	1	0	0	1	0	0	3

Table C.1. AltProts validated by MS searches. (cont.)

#	AltProt Identifier	Type	10-day Nodules	14-day Nodules	28-day Nodules	Buds	Flowers	Leaves	Roots	Seeds	Stems	Whole Plant	Total
54	MtrunA17_Chr4g0065721_3F_162-251_90	m	0	0	0	1	0	1	0	0	0	1	3
55	MtrunA17_Chr5g0396481_1F_676-792_117	m	0	0	0	0	0	1	1	0	1	0	3
56	MtrunA17_Chr5g0404021_1F_1-102_102	m	1	1	0	0	0	0	0	0	0	1	3
57	MtrunA17_Chr5g0438171_2F_350-745_396	m	1	1	0	0	0	0	0	0	0	1	3
58	MtrunA17_Chr5g0448771_1F_1006-1095_90	m	1	1	0	0	1	0	0	0	0	0	3
59	MtrunA17_Chr6g0449601_2F_563-682_120	m	0	0	0	0	1	1	0	0	1	0	3
60	MtrunA17_Chr6g0453531_3F_492-1406_915	m	1	0	0	0	1	0	0	0	0	1	3
61	MtrunA17_Chr6g0454541_1F_274-453_180	m	0	0	0	0	0	0	1	0	1	1	3
62	MtrunA17_Chr6g0457981_3F_123-311_189	m	1	1	0	0	0	0	0	0	0	1	3
63	MtrunA17_Chr6g0488431_1F_2500-2625_126	m	1	0	0	0	1	0	0	1	0	0	3
64	MtrunA17_Chr6g1025074_3F_4749-5060_312	n	0	0	0	1	0	1	0	0	0	1	3
65	MtrunA17_Chr7g0214901_2F_935-1123_189	m	0	0	0	0	0	1	1	0	1	0	3
66	MtrunA17_Chr7g0229931_2F_74-289_216	m	1	1	0	0	0	0	0	0	0	1	3
67	MtrunA17_Chr7g0236061_2F_470-661_192	m	0	1	0	1	0	0	0	0	0	1	3
68	MtrunA17_Chr7g0259571_1F_1684-1770_87	m	0	0	0	1	1	0	0	0	0	1	3
69	MtrunA17_Chr7g0269161_3F_603-692_90	n	1	0	0	1	1	0	0	0	0	0	3
70	MtrunA17_Chr8g0338151_1F_613-678_66	m	0	1	0	0	1	0	0	0	1	0	3
71	MtrunA17_Chr8g0350901_2F_137-346_210	m	1	1	0	0	0	0	0	0	0	1	3
72	MtrunA17_Chr8g0351961_1F_433-510_78	m	0	1	0	0	1	0	0	1	0	0	3
73	MtrunA17_Chr8g0385391_3F_942-1049_108	m	0	0	0	1	0	1	1	0	0	0	3
74	MtrunA17_CPg0493231_3F_3-149_147	m	0	0	0	1	1	0	0	0	0	1	3
75	MtrunA17_Chr1g0149661_1F_3166-3279_114	m	0	0	0	0	0	1	0	0	1	0	2
76	MtrunA17_Chr1g0150541_3F_267-614_348	m	0	0	0	0	0	1	1	0	0	0	2
77	MtrunA17_Chr1g0151431_2F_683-994_312	m	1	0	0	0	1	0	0	0	0	0	2
78	MtrunA17_Chr1g0175451_3F_108-221_114	m	0	0	0	1	1	0	0	0	0	0	2
79	MtrunA17_Chr1g0181271_3F_267-410_144	m	0	1	0	1	0	0	0	0	0	0	2
80	MtrunA17_Chr1g0184331_2F_545-775_231	m	1	1	0	0	0	0	0	0	0	0	2
81	MtrunA17_Chr1g0189931_3F_450-512_63	m	0	0	0	0	0	1	0	1	0	0	2
82	MtrunA17_Chr1g0195991_1F_490-564_75	m	0	1	0	0	0	0	1	0	0	0	2
83	MtrunA17_Chr1g0198561_1F_1120-1191_72	m	0	0	0	0	1	0	0	1	0	0	2
84	MtrunA17_Chr1g0199491_3F_486-650_165	m	0	1	0	0	1	0	0	0	0	0	2
85	MtrunA17_Chr1g0201971_3F_2580-2750_171	m	0	0	0	1	1	0	0	0	0	0	2

Table C.1. AltProts validated by MS searches. (cont.)

#	AltProt Identifier	Type	10-day Nodules	14-day Nodules	28-day Nodules	Buds	Flowers	Leaves	Roots	Seeds	Stems	Whole Plant	Total
86	MtrunA17_Chr1g0207281_1F_1336-1407_72	m	1	1	0	0	0	0	0	0	0	0	2
87	MtrunA17_Chr2g0277311_3F_2238-2480_243	m	0	0	0	0	1	0	0	1	0	0	2
88	MtrunA17_Chr2g0279291_3F_2166-2285_120	m	0	0	0	1	0	0	1	0	0	0	2
89	MtrunA17_Chr2g0280641_3F_666-785_120	m	1	1	0	0	0	0	0	0	0	0	2
90	MtrunA17_Chr2g0285561_2F_200-829_630	m	0	0	0	1	1	0	0	0	0	0	2
91	MtrunA17_Chr2g0287651_1F_1096-1209_114	m	0	0	0	1	1	0	0	0	0	0	2
92	MtrunA17_Chr2g0307381_2F_2-163_162	m	1	1	0	0	0	0	0	0	0	0	2
93	MtrunA17_Chr2g0322121_3F_3342-3554_213	m	1	1	0	0	0	0	0	0	0	0	2
94	MtrunA17_Chr2g0322691_2F_269-373_105	m	0	0	0	0	0	1	0	0	1	0	2
95	MtrunA17_Chr2g1005408_2F_1598-1738_141	n	0	0	0	1	1	0	0	0	0	0	2
96	MtrunA17_Chr3g0083781_1F_82-273_192	m	1	1	0	0	0	0	0	0	0	0	2
97	MtrunA17_Chr3g0084851_2F_3311-3598_288	m	0	0	0	1	0	0	0	0	0	1	2
98	MtrunA17_Chr3g0085021_2F_5-244_240	m	1	1	0	0	0	0	0	0	0	0	2
99	MtrunA17_Chr3g0091141_1F_976-1071_96	m	0	0	0	0	0	0	1	0	1	0	2
100	MtrunA17_Chr3g0106581_1F_2065-2319_255	m	0	0	0	0	0	0	0	1	0	1	2
101	MtrunA17_Chr3g0126961_3F_2757-2849_93	m	0	0	0	1	0	1	0	0	0	0	2
102	MtrunA17_Chr4g0004721_1F_634-813_180	m	1	1	0	0	0	0	0	0	0	0	2
103	MtrunA17_Chr4g0029481_3F_3-203_201	m	1	1	0	0	0	0	0	0	0	0	2
104	MtrunA17_Chr4g0040121_2F_1631-1720_90	m	0	0	0	1	0	0	0	0	1	0	2
105	MtrunA17_Chr4g0047061_3F_156-305_150	m	1	1	0	0	0	0	0	0	0	0	2
106	MtrunA17_Chr4g0059001_2F_83-277_195	m	0	0	0	1	0	0	0	0	1	0	2
107	MtrunA17_Chr4g0065371_2F_143-331_189	n	0	0	0	0	0	0	1	0	1	0	2
108	MtrunA17_Chr4g0069301_2F_1124-1297_174	m	0	0	0	0	0	1	0	0	1	0	2
109	MtrunA17_Chr4g0069311_3F_204-1316_1113	m	0	1	0	0	0	0	0	1	0	0	2
110	MtrunA17_Chr4g0076891_3F_501-641_141	m	0	0	0	0	1	1	0	0	0	0	2
111	MtrunA17_Chr4g1018210_3F_21-332_312	n	0	0	0	0	0	1	0	0	1	0	2
112	MtrunA17_Chr5g0407021_2F_530-673_144	m	0	0	0	0	1	0	0	1	0	0	2
113	MtrunA17_Chr5g0425031_3F_642-773_132	m	0	0	0	0	1	1	0	0	0	0	2
114	MtrunA17_Chr5g0429391_3F_2490-2684_195	m	1	1	0	0	0	0	0	0	0	0	2
115	MtrunA17_Chr5g0429551_2F_404-769_366	m	0	0	0	0	1	0	0	1	0	0	2
116	MtrunA17_Chr5g0432161_2F_1865-2053_189	m	0	0	0	0	0	0	1	0	1	0	2
117	MtrunA17_Chr6g0474051_2F_1061-1228_168	m	1	1	0	0	0	0	0	0	0	0	2

Table C.1. AltProts validated by MS searches. (cont.)

#	AltProt Identifier	Type	10-day Nodules	14-day Nodules	28-day Nodules	Buds	Flowers	Leaves	Roots	Seeds	Stems	Whole Plant	Total
118	MtrunA17_Chr6g0475011_3F_2913-3095_183	m	0	0	0	0	0	0	1	1	0	0	2
119	MtrunA17_Chr6g0476771_3F_333-587_255	m	0	0	0	0	1	0	0	1	0	0	2
120	MtrunA17_Chr6g0478901_3F_1047-1115_69	m	0	0	0	0	0	0	1	0	1	0	2
121	MtrunA17_Chr6g0480011_2F_1244-1336_93	m	0	0	0	0	0	1	0	0	1	0	2
122	MtrunA17_Chr6g0488831_1F_175-525_351	m	1	1	0	0	0	0	0	0	0	0	2
123	MtrunA17_Chr6g1025105_3F_4608-4754_147	n	1	0	0	0	0	0	0	0	0	1	2
124	MtrunA17_Chr7g0214621_3F_159-260_102	m	0	0	0	0	0	1	1	0	0	0	2
125	MtrunA17_Chr7g0217121_2F_647-778_132	m	0	0	0	0	0	1	0	0	1	0	2
126	MtrunA17_Chr7g0226641_2F_563-745_183	m	1	1	0	0	0	0	0	0	0	0	2
127	MtrunA17_Chr7g0226881_1F_826-963_138	m	1	1	0	0	0	0	0	0	0	0	2
128	MtrunA17_Chr7g0237331_3F_2616-2699_84	m	0	0	0	0	0	1	0	0	1	0	2
129	MtrunA17_Chr7g0242451_2F_992-1054_63	m	0	0	0	1	0	0	1	0	0	0	2
130	MtrunA17_Chr7g0248251_3F_1116-1307_192	m	0	1	0	0	0	0	1	0	0	0	2
131	MtrunA17_Chr7g0275421_2F_134-220_87	m	0	0	0	0	1	0	0	1	0	0	2
132	MtrunA17_Chr7g0276191_1F_2128-2388_261	m	0	0	0	0	1	0	0	0	1	0	2
133	MtrunA17_Chr7g1028910_1F_364-480_117	n	0	0	0	0	0	1	0	0	1	0	2
134	MtrunA17_Chr8g0356381_1F_505-630_126	m	0	0	0	1	0	1	0	0	0	0	2
135	MtrunA17_Chr8g0358511_2F_923-985_63	m	0	0	0	0	0	1	0	0	1	0	2
136	MtrunA17_Chr8g0362151_3F_1950-2057_108	m	0	0	0	0	0	1	0	0	1	0	2
137	MtrunA17_Chr8g0377071_2F_1406-1693_288	m	0	0	0	0	0	1	0	0	1	0	2
138	MtrunA17_Chr8g0384631_2F_1109-1363_255	m	0	0	0	0	1	0	0	1	0	0	2
139	MtrunA17_Chr8g0386571_2F_2357-2452_96	m	0	1	0	0	0	0	1	0	0	0	2
140	MtrunA17_Chr8g0388011_1F_412-576_165	m	0	0	0	0	0	0	1	0	1	0	2
141	MtrunA17_Chr8g1039171_1F_3553-3627_75	n	0	0	0	1	1	0	0	0	0	0	2
142	MtrunA17_MTg0491281_3F_729-974_246	n	1	0	0	0	1	0	0	0	0	0	2

APPENDIX D: VALIDATION SUMMARIES OF MS SEARCHES

The following tables, Table D.1, Table D.2, and Table D.3, show the number of validated proteins in MS searches. Note that validated proteins in these tables correspond to identified refProts, altProts, and contaminant sequences. The first column of these tables correspond to the following information: #1: Proteins: Validated; #2: Proteins: Total Possible TP; #3: Proteins: FDR Limit [%]; #4: Proteins: FNR Limit [%]; #5: Proteins: Confidence Limit [%].

Table D.1. Validation summary of the first-step MS searches (mRNA-derived).

#Row	part1-10-day Nodules	part2-10-day Nodules	part3-10-day Nodules	part4-10-day Nodules	part5-10-day Nodules	part6-10-day Nodules	part7-10-day Nodules	part8-10-day Nodules	part9-10-day Nodules	part10-10-day Nodules
#1	9,894	9,954	9,950	9,924	9,919	9,944	9,899	9,934	9,942	9,922
#2	9,966	10,011	9,985	9,982	10,023	10,011	10,024	9,921	9,998	10,009
#3	1	1	1	1	1	1	1	1	1	1
#4	2	2	2	2	2	2	2	2	2	2
#5	60	58	58	57	63	61	65	55	58	65
#Row	**part1-14-day Nodules**	**part2-14-day Nodules**	**part3-14-day Nodules**	**part4-14-day Nodules**	**part5-14-day Nodules**	**part6-14-day Nodules**	**part7-14-day Nodules**	**part8-14-day Nodules**	**part9-14-day Nodules**	**part10-14-day Nodules**
#1	9,825	9,801	9,736	6,480	9,802	8,830	9,766	9,745	9,787	9,829
#2	9,941	9,871	9,894	8,928	9,841	9,354	9,844	9,885	9,921	9,963
#3	1	1	1	1	1	1	1	1	1	1
#4	2	2	3	28	2	7	2	2	3	2
#5	63	59	70	92	57	73	59	60	64	71
#Row	**part1-28-day Nodules**	**part2-28-day Nodules**	**part3-28-day Nodules**	**part4-28-day Nodules**	**part5-28-day Nodules**	**part6-28-day Nodules**	**part7-28-day Nodules**	**part8-28-day Nodules**	**part9-28-day Nodules**	**part10-28-day Nodules**
#1	6,434	6,374	6,323	211	6,357	7,007	6,365	6,301	7,000	6,929
#2	6,930	6,961	6,941	6,289	6,940	7,412	6,957	6,991	7,366	7,317
#3	1	1	1	1	1	1	1	1	1	1
#4	10	11	12	97	11	7	12	12	6	6
#5	67	70	70	92	69	77	68	73	77	83
#6	67	70	70	92	69	77	68	73	77	83

Table D.1. Validation summary of the first-step MS searches (mRNA-derived). (cont.)

#Row	part1-Buds	part2-Buds	part3-Buds	part4-Buds	part5-Buds	part6-Buds	part7-Buds	part8-Buds	part9-Buds	part10-Buds
#1	10,794	10,802	10,837	10,817	10,830	10,881	10,843	10,790	10,807	10,810
#2	10,871	10,858	10,873	10,898	10,875	10,890	10,913	10,859	10,876	10,899
#3	1	1	1	1	1	1	1	1	1	1
#4	2	2	2	2	2	2	2	2	2	2
#5	60	58	56	58	57	56	58	60	58	59
#Row	**part1-Flowers**	**part2-Flowers**	**part3-Flowers**	**part4-Flowers**	**part5-Flowers**	**part6-Flowers**	**part7-Flowers**	**part8-Flowers**	**part9-Flowers**	**part10-Flowers**
#1	12,934	12,942	12,917	12,910	12,915	12,896	12,915	12,962	12,900	12,958
#2	12,875	12,856	12,863	12,843	12,856	12,837	12,840	12,879	12,831	12,833
#3	1	1	1	1	1	1	1	1	1	1
#4	1	1	1	1	1	1	1	1	1	-
#5	51	47	47	48	49	47	48	45	48	42
#6	51	47	47	48	49	47	48	45	48	42
#Row	**part1-Leaves**	**part2-Leaves**	**part3-Leaves**	**part4-Leaves**	**part5-Leaves**	**part6-Leaves**	**part7-Leaves**	**part8-Leaves**	**part9-Leaves**	**part10-Leaves**
#1	8,516	8,783	8,640	8,573	8,700	8,611	8,716	8,624	8,601	8,632
#2	9,389	9,433	9,385	9,383	9,400	9,393	9,419	9,433	9,367	9,379
#3	1	1	1	1	1	1	1	1	1	1
#4	10	8	9	10	8	9	9	10	9	9
#5	87	83	83	85	83	84	85	86	84	85
#Row	**part1-Roots**	**part2-Roots**	**part3-Roots**	**part4-Roots**	**part5-Roots**	**part6-Roots**	**part7-Roots**	**part8-Roots**	**part9-Roots**	**part10-Roots**
#1	6,275	6,253	6,255	6,191	6,205	6,140	6,186	6,271	6,228	6,188
#2	7,606	7,462	7,506	7,544	7,579	7,480	7,528	7,552	7,466	7,541
#3	1	1	1	1	1	1	1	1	1	1
#4	18	17	18	19	19	19	19	18	17	19
#5	93	92	90	91	94	91	92	90	92	90
#Row	**part1-Seeds**	**part2-Seeds**	**part3-Seeds**	**part4-Seeds**	**part5-Seeds**	**part6-Seeds**	**part7-Seeds**	**part8-Seeds**	**part9-Seeds**	**part10-Seeds**
#1	9,875	9,909	9,867	9,866	9,889	9,882	9,882	9,880	9,887	9,888
#2	9,858	9,841	9,822	9,825	9,836	9,859	9,840	9,817	9,845	9,870
#3	1	1	1	1	1	1	1	1	1	1
#4	2	2	2	2	2	2	2	2	2	2
#5	54	50	50	51	49	53	53	48	50	52
#Row	**part1-Stems**	**part2-Stems**	**part3-Stems**	**part4-Stems**	**part5-Stems**	**part6-Stems**	**part7-Stems**	**part8-Stems**	**part9-Stems**	**part10-Stems**
#1	5,571	5,557	5,384	5,361	5,446	5,548	5,545	5,468	5,419	5,608
#2	6,534	6,417	6,428	6,458	6,574	6,370	6,463	6,506	6,454	6,573
#3	1	1	1	1	1	1	1	1	1	1
#4	16	14	17	18	18	14	15	17	17	16
#5	83	84	90	78	84	80	80	88	90	85

Table D.1. Validation summary of the first-step MS searches (mRNA-derived). (cont.)

#Row	part1-Whole Plant	part2-Whole Plant	part3-Whole Plant	part4-Whole Plant	part5-Whole Plant	part6-Whole Plant	part7-Whole Plant	part8-Whole Plant	part9-Whole Plant	part10-Whole Plant
#1	11,550	11,538	11,557	11,550	11,511	11,570	11,593	11,612	11,572	11,544
#2	11,715	11,721	11,714	11,675	11,690	11,709	11,716	11,746	11,755	11,704
#3	1	1	1	1	1	1	1	1	1	1
#4	3	3	2	2	3	2	2	2	3	2
#5	64	68	66	65	68	66	65	67	69	65

Table D.2. Validation summary of the second-step MS searches (mRNA-derived).

#Row	mRNA second step-10-day Nodules	mRNA second step-14-day Nodules	mRNA second step-28-day Nodules	mRNA second step-Buds	mRNA second step-Flowers	mRNA second step-Leaves	mRNA second step-Roots	mRNA second step-Seeds	mRNA second step-Stems	mRNA second step-Whole Plant
#1	10,019	9,776	2,608	10,885	13,026	8,643	6,127	10,008	5,468	11,539
#2	10,024	9,890	6,698	10,959	12,902	9,509	7,567	9,943	6,658	11,787
#3	1	1	1	1	1	1	1	1	1	1
#4	1	3	61	2	0	10	20	2	19	3
#5	69	58	95	61	40	87	93	50	85	65

Table D.3. Validation summary of the second-step MS searches (non-mRNA-derived altProts).

#Row	ncRNA-10-day Nodules	ncRNA-14-day Nodules	ncRNA-28-day Nodules	ncRNA-Buds	ncRNA-Flowers	ncRNA-Leaves	ncRNA-Roots	ncRNA-Seeds	ncRNA-Stems	ncRNA-Whole Plant
#1	9,920	7,520	368	10,802	12,901	8,569	6,200	9,895	5,402	11,578
#2	9,986	9,127	6,418	10,841	12,838	9,390	7,591	9,847	6,529	11,773
#3	1	1	1	1	1	1	1	1	1	1
#4	2	18	94	2	1	10	19	2	18	3
#5	61	91	96	55	47	86	92	53	88	64
#Row	**rRNA-10-day Nodules**	**rRNA-14-day Nodules**	**rRNA-28-day Nodules**	**rRNA-Buds**	**rRNA-Flowers**	**rRNA-Leaves**	**rRNA-Roots**	**rRNA-Seeds**	**rRNA-Stems**	**rRNA-Whole Plant**
#1	9,845	9,726	326	10,768	12,871	8,518	6,101	9,855	5,411	11,453
#2	9,870	9,800	6,296	10,794	12,770	9,313	7,461	9,790	6,387	11,661
#3	1	1	1	1	1	1	1	1	1	1
#4	1	2	95	2	0	9	19	1	16	3
#5	70	56	96	57	43	86	93	47	78	69
#Row	**tRNA-10-day Nodules**	**tRNA-14-day Nodules**	**tRNA-28-day Nodules**	**tRNA-Buds**	**tRNA-Flowers**	**tRNA-Leaves**	**tRNA-Roots**	**tRNA-Seeds**	**tRNA-Stems**	**tRNA-Whole Plant**
#1	9,878	9,656	336	10,767	12,859	8,484	6,119	9,862	5,579	11,472
#2	9,872	9,735	6,262	10,788	12,740	9,315	7,478	9,803	6,405	11,664
#3	1	1	1	1	1	1	1	1	1	1
#4	1	2	95	2	0	10	19	1	14	3
#5	62	59	96	56	40	88	94	49	89	67

APPENDIX E: RELATIVE POSITION OF THE FIRST IN-FRAME START CODON AUG PER RNA TYPE

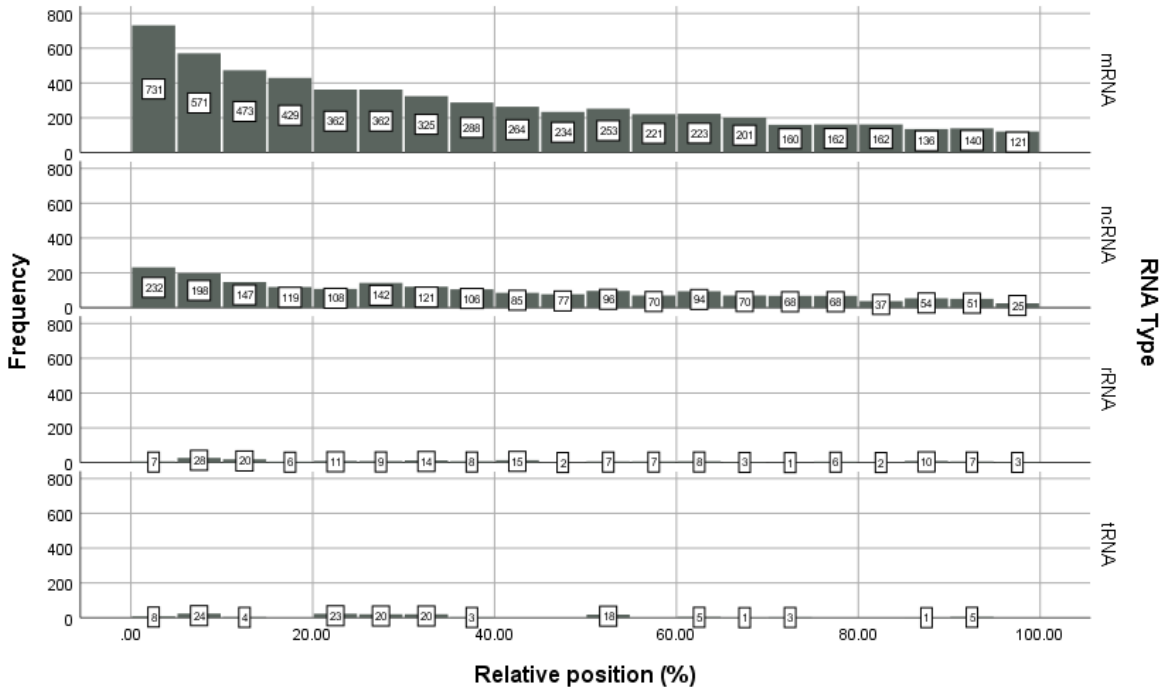

Figure E.1. Relative position of the first in-frame start codon AUG in different transcript types. mRNA: $\bar{x}$ = 35.7, SD = 27.8, n = 5,818; ncRNA: $\bar{x}$ = 36.5, SD = 27.5, n = 1,968; rRNA: $\bar{x}$ = 37.2, SD = 28.2, n = 174; tRNA: $\bar{x}$ = 30.1, SD = 22.1, n = 135.

APPENDIX F: ILLUSTRATION OF RIBOSOMAL FRAMESHIFTING SITES FOR MTRUNA17_CHR1G0156271 AND MTRUNA17_CHR1G0185811

Ribosomal frameshifting positions were illustrated for MtrunA17_Chr1g0156271 and MtrunA17_Chr1g0185811 in the following figures.

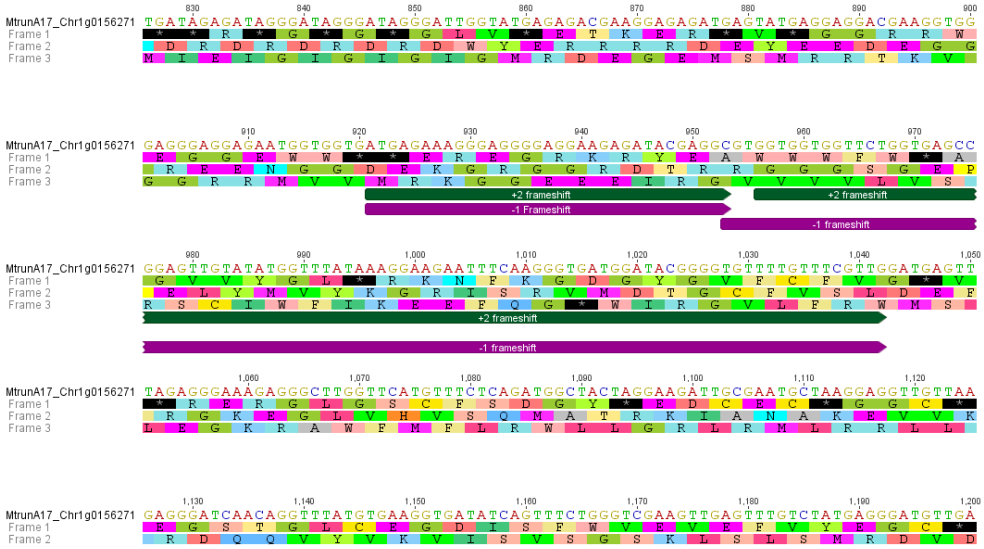

Figure F.1. Ribosomal frameshifting events on MtrunA17_Chr1g0156271

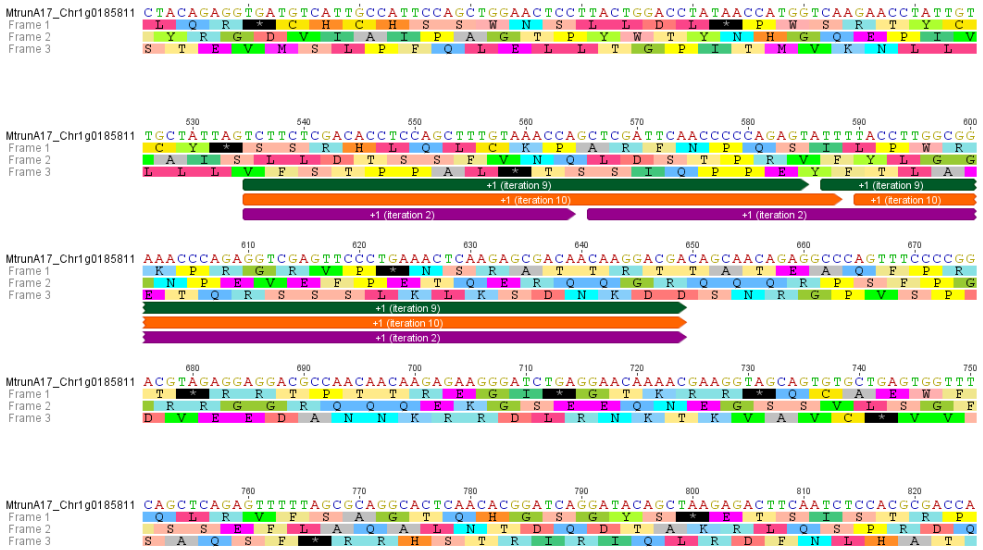

Figure F.2. Ribosomal frameshifting events on MtrunA17_Chr1g0185811

APPENDIX G: LITERATURE SEARCH FOR GENES THAT WERE VALIDATED BY ALTORF TRANSLATION

Comprehensive literature searches for genes of altORF that were validated in this study is shown in the table. References of the studies and altProt IDs are available in the supplementary file; see section 2.7 Data Availability.

Table G.1. Comprehensive literature searches for genes of validated altORFs.

	Gene Symbol	AltProt ID	Transcript Type	Biological Process	AltProt Type/Comments	% Identity of the Top Hit	Number of Hits
1	MtPIN3	MtrunA17_Chr1g0160461 _1F_1945-2175_231	mRNA	SNF	Conserved	45.3	1
2	MtSYT3	MtrunA17_Chr1g0199571 _2F_863-1120_258	mRNA	SNF	Conserved	46.6	1
3	MtARF3	MtrunA17_Chr2g0282961 _1F_1447-1602_156	mRNA	SNF	Conserved	83.3	2
4	MtCYP72A67	MtrunA17_Chr2g0288661 _2F_392-592_201	mRNA	SNF; saponin metabolism	Conserved	64.1	1
5	MtVAMP721d	MtrunA17_Chr2g0291651 _1F_718-990_273	mRNA	SNF; AM symbiosis	Conserved	64.6	3
6	MtCNGC15c	MtrunA17_Chr2g0326871 _3F_1074-1199_126	mRNA	SNF; AM symbiosis	Conserved	77.5	3
7	MtPLT1	MtrunA17_Chr2g0328971 _2F_866-994_129	mRNA	SNF	Conserved	67.5	1
8	MtYSL7	MtrunA17_Chr3g0109311 _1F_1357-1974_618	mRNA	SNF	Conserved	40.5	1
9	MtGS1b	MtrunA17_Chr3g0110261 _2F_308-784_477	mRNA	SNF	Conserved	57.8	2
10	MtNRAMP1	MtrunA17_Chr3g0124971 _3F_489-662_174	mRNA	SNF	Conserved	58.6	2
11	MtNRAMP1	MtrunA17_Chr3g0124971 _3F_1005-1178_174	mRNA	SNF	Conserved	93.1	2
12	MtYSL3	MtrunA17_Chr3g0127441 _1F_991-1137_147	mRNA	SNF	Conserved	64.1	1
13	MtKNOX5	MtrunA17_Chr3g0137241 _1F_1387-1485_99	mRNA	SNF	Conserved; no overlapping protein-coding gene known for this locus	100	1
14	MtABCG59	MtrunA17_Chr3g0138261 _1F_2491-2601_111	mRNA	SNF; AM symbiosis	Conserved	67.6	1
15	MtABCG59	MtrunA17_Chr3g0138261 _2F_3491-3688_198	mRNA	SNF; AM symbiosis	Conserved	56.1	3
16	MtNLP1	MtrunA17_Chr3g0143921 _1F_3046-3132_87	mRNA	SNF	Conserved	93.1	3
17	MtGbeta1	MtrunA17_Chr3g0144511	mRNA	SNF	Conserved	93.5	1

Table G.1. Comprehensive literature searches for genes of validated altORFs. (cont.)

	Gene Symbol	AltProt ID	Transcript Type	Biological Process	AltProt Type/Comments	% Identity of the Top Hit	Number of Hits
		_1F_778-1101_324					
18	MtGbeta1	MtrunA17_Chr3g0144511_3F_1344-1475_132	mRNA	SNF	Conserved	82.4	1
19	MtFPN2	MtrunA17_Chr4g0004871_1F_1252-1470_219	mRNA	SNF	Conserved	55.3	1
20	MtP5CS3	MtrunA17_Chr4g0008951_1F_1675-1812_138	mRNA	SNF; salt stress; drought	Conserved	87	1
21	MtPHO2-like	MtrunA17_Chr4g0009054_1F_1300-1440_141	mRNA	SNF	Conserved	95.3	95
22	MtPHO2-like	MtrunA17_Chr4g0009054_1F_3370-3927_558	mRNA	SNF	Conserved; no overlapping protein-coding gene known for this locus	100	14,113
23	MtPHO2-like	MtrunA17_Chr4g0009054_1F_4195-4350_156	mRNA	SNF	Conserved	83.7	189
24	MtPHO2-like	MtrunA17_Chr4g0009054_2F_4052-4372_321	mRNA	SNF	Conserved	98.8	41
25	MtPHO2-like	MtrunA17_Chr4g0009054_3F_1380-2345_966	mRNA	SNF	Conserved	95.1	608
26	MtCNGC15b	MtrunA17_Chr4g0028861_1F_997-1134_138	mRNA	SNF; AM symbiosis	Conserved	81	4
27	MtRab7a2	MtrunA17_Chr4g0034871_1F_1102-1335_234	mRNA	SNF	Conserved	44.6	1
28	MtVAMP721e	MtrunA17_Chr4g0043521_2F_647-790_144	mRNA	SNF; AM symbiosis	Conserved	76.9	3
29	MtRIT	MtrunA17_Chr4g0043744_3F_2244-2471_228	mRNA	SNF	Conserved	88.7	2
30	MtAKT1	MtrunA17_Chr4g0063141_3F_528-854_327	mRNA	SNF	Conserved	60.9	2
31	MtSUCS1	MtrunA17_Chr4g0070011_3F_1743-1832_90	mRNA	SNF	Conserved	85.2	1
32	MtSUCS1	MtrunA17_Chr4g0070011_3F_1743-1832_90_MtrunA17_Chr4g0070011_1F_58-2481_2424_-1_iteration_16_Within_5'_of_altprot_12584_Chimeric	mRNA	SNF	MS-supported chimeric	NA	NA
33	MtHAN1	MtrunA17_Chr5g0404131_3F_612-854_243	mRNA	SNF	Conserved	75	1
34	MtLYK3	MtrunA17_Chr5g0439631_3F_2253-2549_297	mRNA	SNF	Conserved; no overlapping protein-coding gene known for this locus	100	644
35	MtCBS1	MtrunA17_Chr6g0469911_1F_1768-1959_192	mRNA	SNF	Conserved	75	1
36	MtCBS1	MtrunA17_Chr6g0469911_3F_1719-1925_207	mRNA	SNF	Conserved	97.6	3
37	MtPIN4	MtrunA17_Chr6g0478431_1F_1735-2004_270	mRNA	SNF	Conserved	63.5	1
38	MtGS1a	MtrunA17_Chr6g0479141_3F_720-875_156	mRNA	SNF	Conserved	54	1

Table G.1. Comprehensive literature searches for genes of validated altORFs. (cont.)

	Gene Symbol	AltProt ID	Transcript Type	Biological Process	AltProt Type/Comments	% Identity of the Top Hit	Number of Hits
39	MtCAS31	MtrunA17_Chr6g0484671 _3F_786-1043_258	mRNA	SNF	MS-supported	NA	NA
40	MtP5CS2	MtrunA17_Chr7g0239721 _2F_1748-1957_210	mRNA	SNF; salt stress; drought	Conserved	71.7	1
41	MtLAX2	MtrunA17_Chr7g0241841 _1F_940-1320_381	mRNA	SNF	Conserved	30.8	1
42	MtSCR	MtrunA17_Chr7g0245601 _2F_2153-2392_240	mRNA	SNF; root development; shoot development	MS-supported	NA	NA
43	MtABCG56	MtrunA17_Chr7g0261971 _2F_4133-4396_264	mRNA	SNF	Conserved	31.3	1
44	MtMCA8	MtrunA17_Chr7g0263361 _3F_1311-1586_276	mRNA	SNF; AM symbiosis	Conserved	51.6	1
45	MthGSHSb	MtrunA17_Chr7g0273141 _1F_625-816_192	mRNA	SNF	MS-supported conserved	91.7	255
46	MthGSHSb	MtrunA17_Chr7g0273141 _2F_2-136_135	mRNA	SNF	Conserved	93.1	26
47	MtNSP1	MtrunA17_Chr8g0344101 _3F_597-752_156	mRNA	SNF; AM symbiosis	Conserved	85.4	1
48	MtMATE67	MtrunA17_Chr8g0352151 _1F_1717-1959_243	mRNA	SNF	Conserved	50	1
49	MtARP3	MtrunA17_Chr8g0381261 _3F_1431-1625_195	mRNA	SNF	Conserved	74.4	1
50	MtSymREM1	MtrunA17_Chr8g0386521 _1F_2020-2235_216	mRNA	SNF	Conserved	56	2
51	MtNCR055	MtrunA17_Chr1g0166851 _1F_1198-1323_126	ncRNA	SNF	Conserved; not MtNCR055	95.2	2
52	MtNCR055	MtrunA17_Chr1g0166851 _3F_1383-1538_156	ncRNA	SNF	Conserved; not MtNCR055	75	1
53	MtCLE34	MtrunA17_Chr2g0325371 _2F_917-1012_96	ncRNA	SNF	Conserved; not MtCLE34; no overlapping protein-coding gene known for this locus	100	1
54	MtNCR035	MtrunA17_Chr4g0007841 _2F_1223-1384_162	ncRNA	SNF	Conserved; not MtNCR035	65.3	40
55	MtNCR211	MtrunA17_Chr4g0018031 _2F_179-358_180	ncRNA	SNF	Conserved; not MtNCR211	77.4	18
56	MtNCR211	MtrunA17_Chr4g0018031 _3F_249-344_96	ncRNA	SNF	Conserved; MtNCR211	100	1
57	MtNCR247	MtrunA17_Chr5g0423671 _1F_907-1068_162	ncRNA	SNF	Conserved; not MtNCR247	60.8	4
58	MtNCR247	MtrunA17_Chr5g0423671 _2F_944-1078_135	ncRNA	SNF	Conserved; not MtNCR247	66.7	1
59	MtNCR247	MtrunA17_Chr5g0423671 _3F_819-1118_300	ncRNA	SNF	Conserved; not MtNCR247	72.7	21
60	MtNCR044	MtrunA17_Chr7g0216231 _3F_1035-1166_132	ncRNA	SNF	Conserved; not MtNCR044	90.3	1
61	MtNCR169	MtrunA17_Chr7g0229931 _2F_74-289_216	ncRNA	SNF	MS-supported conserved; MtNCR169	100	1
62	MtENOD40-1	MtrunA17_Chr8g0368441	ncRNA	SNF	Conserved; not	100	1

Table G.1. Comprehensive literature searches for genes of validated altORFs. (cont.)

	Gene Symbol	AltProt ID	Transcript Type	Biological Process	AltProt Type/Comments	% Identity of the Top Hit	Number of Hits
		_1F_205-414_210			MtENOD40-1; this locus overlaps with a protein-coding gene MtrunA17_Chr 8g0368434		
63	MtSERF1	MtrunA17_Chr1g0170481 _1F_367-609_243	mRNA	Embryogenesis	Conserved	67.4	2
64	MtAGa	MtrunA17_Chr2g0284911 _2F_314-505_192	mRNA	Flower development	Conserved	64	1
65	MtREV1	MtrunA17_Chr2g0326731 _1F_958-1023_66	mRNA	Leaf development	MS-supported	NA	NA
66	MtREV1	MtrunA17_Chr2g0326731 _2F_1352-1528_177	mRNA	Leaf development	Conserved	88	1
67	MtMATE66	MtrunA17_Chr2g0328761 _2F_1205-1498_294	mRNA	Al3+ tolerance; Fe homeostasis	Conserved	69.4	1
68	MtMATE66	MtrunA17_Chr2g0328761 _3F_1611-1745_135	mRNA	Al3+ tolerance; Fe homeostasis	Conserved	53.3	1
69	MtCCR1	MtrunA17_Chr2g0333781 _3F_1632-1784_153	mRNA	Lignin metabolism; stem, leaf, and flower development	Conserved	66.7	1
70	MtMATE1	MtrunA17_Chr5g0442331 _2F_1610-1888_279	mRNA	Flavonoid metabolism; seed composition	Conserved; no overlapping protein-coding gene known for this locus	100	1
71	MtPIN10	MtrunA17_Chr7g0255941 _2F_1817-2038_222	mRNA	Leaf development; cotyledon development; flower development	Conserved	50	1
72	MtDefMd1	MtrunA17_Chr8g0339711 _3F_363-524_162	mRNA	AM symbiosis	Conserved; no overlapping protein-coding gene known for this locus	100	151
73	MtAGb	MtrunA17_Chr8g0380021 _2F_659-850_192	mRNA	Flower development	Conserved	72	1
74	MtLHA	MtrunA17_Chr8g0388921 _1F_1267-1512_246	mRNA	Saponin metabolism	Conserved	54	1
75	MtSTF	MtrunA17_Chr8g0392991 _2F_530-736_207	mRNA	Leaf development	Conserved	64.4	1