UNIVERSIDADE ESTADUAL PAULISTA - UNESP

CÂMPUS DE JABOTICABAL

IDENTIFICATION OF NOVEL CODING SEQUENCES ASSOCIATED WITH SEASONALITY AND WATER RESTRICTION IMPOSITION IN Eucalyptus grandis BY MASS SPECTROMETRY

Gabriel Lemes Jorge

Agronomist

2022

UNIVERSIDADE ESTADUAL PAULISTA - UNESP

CÂMPUS DE JABOTICABAL

IDENTIFICATION OF NOVEL CODING SEQUENCES ASSOCIATED WITH SEASONALITY AND WATER RESTRICTION IMPOSITION IN Eucalyptus grandis BY MASS SPECTROMETRY

Gabriel Lemes Jorge

Advisor: Prof. Dr. Tiago Santana Balbuena

Tese apresentada à Faculdade de Ciências Agrárias e Veterinárias - Unesp, Câmpus de Jaboticabal, como parte das exigências para a obtenção do título de Doutor em Agronomia (Genética e Melhoramento de Plantas)

2022

J82i

Jorge, Gabriel Lemes
Identification of novel coding sequences associated with seasonality and water restriction imposition in Eucalyptus grandis by mass spectrometry / Gabriel Lemes Jorge. -- Jaboticabal, 2022
115 p. : il., tabs.

Tese (doutorado) - Universidade Estadual Paulista (Unesp), Faculdade de Ciências Agrárias e Veterinárias, Jaboticabal
Orientador: Tiago Santana Balbuena

1. Proteomics. 2. Proteogenomics. 3. drought stress. 4. PRM. 5. SRM. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Biblioteca da Faculdade de Ciências Agrárias e Veterinárias, Jaboticabal. Dados fornecidos pelo autor(a).

Essa ficha não pode ser modificada.

UNIVERSIDADE ESTADUAL PAULISTA

Câmpus de Jaboticabal

CERTIFICADO DE APROVAÇÃO

TÍTULO DA TESE: IDENTIFICATION OF NOVEL CODING SEQUENCES ASSOCIATED WITH SEASONALITY AND WATER RESTRICTION IMPOSITION IN *Eucalyptus grandis* BY MASS SPECTROMETRY

AUTOR: GABRIEL LEMES JORGE
ORIENTADOR: TIAGO SANTANA BALBUENA

Aprovado como parte das exigências para obtenção do Título de Doutor em AGRONOMIA (GENÉTICA E MELHORAMENTO DE PLANTAS), pela Comissão Examinadora:

Pesquisador Dr. TIAGO SANTANA BALBUENA (Participaçao Virtual)
Departamento de Tecnologia / FCAV UNESP Jaboticabal

Profa. Dra. MARIA HELENA DE SOUZA GOLDMAN (Participaçao Virtual)
Departamento de Biologia-FFCLRP/USP / Ribeirão Preto/SP

Prof. Dr. DANIEL GUARIZ PINHEIRO (Participaçao Virtual)
Departamento de Biotecnologia Agropecuária e Ambiental / FCAV / UNESP - Jaboticabal

Prof. Dr. FÁBIO CÉSAR SOUSA NOGUEIRA (Participaçao Virtual)
UFRJ / Rio de Janeiro, RJ

Prof. Dr. JESUS APARECIDO FERRO (Participaçao Virtual)
Departamento de Biotecnologia Agropecuária e Ambiental / UNESP - Jaboticabal

Jaboticabal, 21 de fevereiro de 2022

AUTHOR'S INFORMATION

Gabriel Lemes Jorge, born on July 27th, 1992, in Araguari, MG, Brazil, has a bachelor's degree in Agronomy from the Federal University of Uberlandia (2016). During his undergraduate, he was a science without borders scholarship holder sponsored by CAPES (2013-2014) and did his internship at the Plant Productivity Laboratory at Trent University, Peterborough, Canada. It went so well that he later came back as a visiting research scholar during his master's sponsored by the Emerging Leaders in the Americas Program (ELAP) that also took place at Trent University, Canada. In his master's dissertation, he was dedicated to studying the selection of soybean lines under biotic stresses and endophytic bacteria as an alleviating factor of drought during lentil cultivation. Gabriel got his master's degree in Agronomy (Crop Science) at the Federal University of Uberlandia in 2018. He started his Ph.D. in the same year at the Sao Paulo State University, Jaboticabal campus, in Agronomy (genetics and plant breeding). In his Ph.D. project, he uses different proteogenomics approaches to identify novel coding sequences in the *Eucalyptus grandis* proteome. Despite all challenges imposed by COVID-19, he went for a 7-month research period as a visiting research scholar sponsored by CNPq at the University of Columbia, Missouri, in the United States focusing on targeted proteomics experiments (2021-2022).

EPIGRAPH

There are three paths to failure: not teaching what you know, not practicing what you teach, and not asking what you don't know (Saint Bede, the Venerable). However, there are three paths to success: teaching what you know, that is, mental generosity, practicing what you teach, that is, ethical coherence, and asking what you ignore, that is, intellectual humility (Mario Sergio Cortella).

I dedicate,

To my parents, Cesar Jorge and Luciane.

To my brother, Vinicius.

ACKNOWLEDGEMENTS

I thank God with all my faith and trust.

I thank my parents, Cesar and Luciane, and my brother, Vinicius, for encouraging, supporting, and giving me the best example to follow that I could ever get from a family. I extend my thanks to my aunt, Viviane. Their support got me through the roughest times, and for that, I can never thank them enough.

My supervisor, Tiago Santana Balbuena that believed in me since the first moment I met him even though I did not know much about Proteomics back then. He inspired me to learn as much as I could and to always give my best. I am grateful for his knowledge, guidance, and patience in our discussions helping me to consolidate and write this thesis.

Special thanks to prof. Jay Thelen, for receiving me as a Research Scholar at his lab at the University of Missouri and providing me all support and guidance.

Major thanks to Prof. Dr. Rinaldo Cesar de Paula, Prof. Dr. Dong Xu, Dr. Jiang Yuexu, Dr. Kim Daewon, and Dr. José Roberto da Silva Nascimento for all help and valuable discussions.

To my Brazilian lab mates, Amanda, Felipe, Patrícia, and Naiara, who were essential in many parts of the conduction, discussion, and execution of the experiments.

Thanks to the National Council for Scientific and Technological Development (CNPq), Brazil, and Sao Paulo Research Foundation (FAPESP), Brazil for the research fellowship. Also, thanks CNPq for the Ph.D. scholarship and sandwich doctoral scholarship.

Thanks to all my family members. They mean the world to me. I cannot cite everyone but a special thanks to my grandparents, Teresinha and Wilson, who are examples of strength, determination, and courage to me. Also, thanks for your support, Cristiane, Adriano, Vitória, and Letícia.

And to those people who appear in our lives, and not by chance, remain, you, my friends, were fundamental in this accomplishment.

A special thanks to Colégio Berlaar Sagrado Coração de Jesus (located in Araguari, MG) and all its team for teaching me, from childhood to the end of high

school, essential values that allow me to lead life with integrity and honesty. In particular, I would like to thank again my high school basketball coach Rita de Cássia (a previous acknowledgment was given in my master's thesis), for teaching me that it is not possible to reach the top of a ladder without having climbed it step by step.

Thanks to the defense committee for accepting the invitation and adding valuable considerations for this thesis.

And last but not least, I thank The School of Agricultural and Veterinary Studies (FCAV) of Unesp, Jaboticabal campus, and the Agronomy Graduate Program on Genetics and Plant Breeding.

TABLE OF CONTENTS

IDENTIFICATION OF NOVEL CODING SEQUENCES ASSOCIATED WITH SEASONALITY AND WATER RESTRICTION IMPOSITION IN *Eucalyptus grandis* BY MASS SPECTROMETRY

ABSTRACT - Understanding how plants cope with current climate change and escalating frequency of drought events is crucial for agriculture. Among the plant species of economic importance, *Eucalyptus* species are widely cultivated for their several industrial applications. However, the current negative effects of changing climate may impose additional challenges on *Eucalyptus* plant adaptability and development. In this scenario, the study of proteins can provide essential insights into how plants regulate many biological processes under such conditions. Here, we use proteomics, a cutting-edge technology that uses mass spectrometry to identify proteins by correlating known virtual spectra against experimental spectra acquired from large-scale analyses of protein mixtures. We also use a proteogenomics approach that can identify novel peptides not initially predicted by conventional proteomics workflows in the *Eucalyptus grandis* proteome. First, our goal was to identify and validate novel peptide forms from *E. grandis* stem samples using a dedicated proteogenomics workflow, which consisted of spectral correlations against modified RNA databases and a de novo peptide sequencing approach. Later, this strategy was applied on a broader scale to identify novel peptide forms in *E. grandis* leaf samples subjected to the effect of seasonality and water restriction imposition. Known protein and novel peptide identifications provided important insights on how the leaf proteome profile of *E. grandis* trees changed over different seasons of the year, as well as under water restriction imposition. Those novel peptide identifications, mainly related to the photosynthesis process, decreased the protein stability by altering the quantitative change upon ΔΔG values and non-covalent interactions. Validation of novel identifications was carried out using either Multiple Reaction Monitoring (MRM) or Parallel Reaction Monitoring (PRM) targeted assays that provided further mass spectrometry support for the existence of the novel peptide identifications.

Keywords: Proteomics, Proteogenomics, drought stress, PRM, SRM.

IDENTIFICAÇÃO DE NOVAS SEQUÊNCIAS CODIFICADORAS ASSOCIADAS À SAZONALIDADE E RESTRIÇÃO HÍDRICA EM *Eucalyptus grandis* POR ESPECTROMETRIA DE MASSAS

RESUMO - Compreender como as plantas lidam com as atuais mudanças climáticas e a crescente frequência de eventos de seca é crucial para a agricultura. Dentre as espécies vegetais de importância econômica, espécies de *Eucalyptus* são amplamente cultivadas por suas diversas aplicações industriais. No entanto, os atuais efeitos negativos das mudanças climáticas podem impor desafios adicionais à adaptabilidade e desenvolvimento de plantas de eucalipto. Nesse cenário, o estudo de proteínas pode fornecer insights essenciais sobre como as plantas regulam muitos processos biológicos sob tais condições. Aqui, usamos a proteômica, uma tecnologia de ponta que usa espectrometria de massa para identificar proteínas por meio da correlação entre espectros virtuais conhecidos e espectros experimentais adquiridos a partir de análises em larga escala de amostras proteicas. Também usamos uma abordagem de proteogenômica que pode identificar novos peptídeos não previstos inicialmente por fluxos de trabalho de proteômica convencionais no proteoma de *Eucalyptus grandis*. Inicialmente, nosso objetivo foi identificar e validar novas sequências de peptídeos em amostras caulinares de *E. grandis* usando um fluxo de trabalho de proteogenômica dedicado que consiste em correlações espectrais contra bancos de dados modificados de RNA e uma abordagem de sequenciamento *de novo*. Posteriormente, essa estratégia foi aplicada em larga escala para a identificação de novas formas peptídicas em amostras de folhas de *E. grandis* submetidas ao efeito da sazonalidade e da imposição de restrição hídrica. Proteínas conhecidas e novas identificações de peptídeos forneceram importantes insights sobre a alteração do proteoma foliar de árvores de *E. grandis* ao longo de diferentes estações do ano, bem como sob a imposição de restrição hídrica. Essas identificações de peptídeos novos, principalmente relacionadas ao processo de fotossíntese, diminuíram a estabilidade da proteína por alterar valores quantitativos de ΔΔG e interações não-covalentes. A validação de novas identificações peptídicas foi realizada por ensaios direcionados de Monitoramento de Reação Múltipla (MRM) ou Monitoramento de Reação Paralela (PRM) que forneceram suporte adicional de espectrometria de massa para a existência das novas sequências peptídicas identificadas.

Palavras-chave: Proteômica, Proteogenômica, estresse hídrico, PRM, SRM.

CHAPTER 1 – General considerations

1. INTRODUCTION

Climate change and alterations in rainfall patterns may impose additional challenges for plant adaptation and development (Easterling *et al.*, 2000). Global warming is projected to increase the frequency and intensity of drought events, which may lead to negative impacts on agriculture and crops of economic importance (Li *et al.*, 2020b). Among them, *Eucalyptus* species stand out for their wide application in the industry as raw material for paper and cellulose production, charcoal, essential oil, and woodwork products. Therefore, changing climate and drought events are more likely to negatively impact *Eucalyptus species* plantations.

In this scenario, proteomics is an essential tool for understanding biological processes from the analysis of the protein profile of a given organism under a given condition since it provides direct information on the end products of gene expression: the proteins (Feist and Hummon, 2015). Furthermore, beyond identifying known proteins, the identification of novel peptide forms is indeed crucial since conventional proteomics workflows may lack the capacity to identify many peptides of interest due to database restrictions (Nesvizhskii, 2014). Proteogenomics, is an alternative approach that provides an extra level of identification (Nesvizhskii, 2014), which could assist in understanding the molecular mechanisms of plant adaptation to environmental changes as novel peptide forms are discovered.

Novel peptide forms can be identified following several pipelines depending on the goal of the research. Here, we used two different approaches, which were based on spectral correlations of experimental data against modified RNA databases (translated into alternative reading frames) and a dedicated *de novo* peptide sequencing workflow. Novel peptide forms with single amino acid substitution (SAS) peptides can often provide novel insights towards better understanding plants' molecular events and, for instance, have a significant impact on protein stability (Teng *et al.*, 2010).

Beyond identifying novel peptide forms, targeted proteomics experiments can provide an extra level of confidence for novel identifications (Gibbons *et al.*, 2019). SRM (Selected Reaction Monitoring) or MRM (Multiple Reaction Monitoring) targeted

experiments are commonly used as they offer more reproducible, selective, and specific results (Percy *et al.*, 2013; Gao *et al.*, 2018). Those mass spectrometric based approaches can work as fundamental methods for data validation, and depending on the goal of the research, increasing the quantification sensitivity of mass spectrometry experiments (Shi *et al.*, 2016).

Here a conventional proteomics workflow and a dedicated proteogenomics approach were used to contribute to the characterization of the proteome of *Eucalyptus grandis* through a post-genomic approach. We also aimed to: 1) confidently identify known proteins and novel peptides forms in the stem and leaf tissues from *E. grandis* plants; 2) provide mapping and genomic characterization of novel peptide forms; 3) investigate how known proteins and novel peptides behave under the effect of seasonality and water restriction imposition over field conditions; 4) deliver insights into protein stability associated with novel peptide identifications under seasonality and water restriction imposition; and 5) validate novel peptide identifications by targeted proteomics assays.

2. LITERATURE REVIEW

2.1. Brief history of *Eucalyptus* genus in Brazil

The *Eucalyptus* genus is important by its wide range of uses and high adaptation to different types of climates and regions. The history behind its first introduction in Brazil is difficult to be precisely defined due to the lack of documents of possible entry dates of seeds and seedlings in the country (Castro *et al.*, 2016). It is believed that the first *Eucalyptus* seedlings were planted around 1824 at the Rio de Janeiro Botanical Garden in Brazil (Marchiori, 2014). Actual studies with the *Eucalyptus* genus began in 1904 by Edmundo Navarro de Andrade (in the Sao Paulo Railroad Company), who introduced a variety of species of different origins within the genus in Jundiaí Garden, Sao Paulo (Castro *et al.*, 2016). The first *Eucalyptus* usages were destined to attend the wood demand for railways' construction (SAMPAIO, 1961). The main interest emerged as for the easy adaptation of *Eucalyptus* species to different environmental conditions in Brazil, which were similar to the Australian latitude, and around 1941, the first breeding attempts were started by Edmundo Navarro de Andrade and Carlos Arnaldo Krug (from the Agronomic Institute of Campinas (IAC) (Castro *et al.*, 2016). Currently, this tree genus is considered one of the most important for the country and other tropical and subtropical regions of the world (IBÁ - Indústria Brasileira de Árvores, 2020).

2.2. *Eucalyptus* economic importance

According to IBÁ - Indústria Brasileira de Árvores (2020), the states of Minas Gerais, Sao Paulo, Mato Grosso do Sul, Paraná, Rio Grande do Sul, and Santa Catarina are the ones with the largest planted forests in Brazil, respectively. Also, during 2019, the total area of planted trees totalized 9.0 million hectares, a 2.4% increase over 2018 (8.79 million hectares). Out of this total, the majority (77%) is represented by *Eucalyptus* trees, with 6.97 million hectares, whereas 18% is represented by pine forests, with 1.64 million hectares. Furthermore, there are an additional 0.39 million hectares planted with other species, such as rubber, acacia, teak, and parica trees. The average yield for *Eucalyptus* plantations in Brazil was 35.3

m^3 ha^{-1} per year in 2019. Interestingly, there are more than 1000 cities where the forest industry works in Brazil, and its gross revenue in 2019 reached R$ 97.4 billion. Lastly, and still according to IBÁ - Indústria Brasileira de Árvores (2020), the ranking for the share of three activities in comprising the gross domestic product of 2019 was: 1) pulp, paper, and paper product manufacturing (46.5%); 2) forest production (wood charcoal, wood (logs and by-products), firewood) (36.2%); and 3) wood product manufacturing (e.g., wood laminates, plywood, pressboard, and particleboard) (17.3%).

2.3. Climate change and drought impact on plants

Climate change is one of the utmost challenges that agriculture faces. Main climate-change-related impacts on biological systems are predicted to occur, and it may outpace intrinsic adaptive capacities by increasing the relative vulnerabilities of many organisms (Staudinger *et al.*, 2013). Global warming and climate change may increase the effects of abiotic stresses on crop production since the climate continues to change (Li et al., 2020). Global climate change is projected to increase the frequency of drought occurring under warming temperatures (Easterling et al., 2000).

Water deficit is an inevitable consequence of climate change and has a substantial impact on terrestrial plant development. The sensitivity of the plant growth stage to water deficit can be affected by many factors, including climatic conditions, crop species, cultivars, agronomic management practices, among others (Chai *et al.*, 2016). As a result, water restriction can cause diverse modifications at the molecular, cellular, and physiological levels, delaying plant growth (Jiménez *et al.*, 2013; Ngara *et al.*, 2018).

Environmental stresses trigger an extensive variety of plant responses, which include altered gene expression and cellular metabolism to changes in growth rate and plant yield (Shao *et al.*, 2009). Those plant reactions exist to avoid the harmful effects triggered by a wide range of both abiotic and biotic stresses, such as drought, light, salinity, and high temperatures (Shao *et al.*, 2009). Among the environmental stresses, drought is one of the most hostile factors for plant growth and productivity, and plant drought stress tolerance may be attributed to molecular responses (e.g., expression of stress-responsive genes), biochemical responses (e.g., decreased efficiency of

Rubisco and transient decrease in photochemical efficiency), and physiological responses (e.g., reduced photosynthesis rate) (Shao *et al.*, 2009).

Short-term water stress experiments applied on *Eucalyptus* species (in Australia) and *Quercus* species (in Spain) revealed adaptive interspecific differences in drought responses between both experiments indicating that drought imposes limitations on Rubisco activity and RuBP regeneration capacity alongside declines in stomatal and mesophyll conductance (Zhou *et al.*, 2014). Also, gene expression in two contrasting hybrid clones of *Eucalyptus camaldulensis x Eucalyptus urophylla* grown under water deficit conditions pointed out significant impacts on photosynthesis, respiration, carbohydrate metabolism, and ion absorption (Martins *et al.*, 2018).

2.4. Conventional Proteomics workflows

Proteins regulate many cellular functions and are essential cellular machinery that enables different tasks within a given biological condition (Feist and Hummon, 2015). Proteomics is an essential approach as it provides direct information on the end products of gene expression: proteins. Thus, proteomics is a technique that seeks to understand biological processes from the analysis of the protein profile of a given organism under a given condition. The field of proteomics is the large-scale study of proteins and relies on different methodologies for protein identification, such as mass spectrometry-based approaches. Two main approaches for mass spectrometry-based proteomics can be cited as top-down and bottom-up analyses. The top-down analysis aims to study whole proteins, whereas bottom-up seeks to investigate peptides from digested proteins. In the bottom-up workflows, proteins are extracted from biological tissue, quantified, in-gel, or in-solution digested, and are finally ready for liquid chromatography-mass spectrometry (LC-MS/MS). One of the acquisition methods that can be used is data-dependent acquisition (DDA) in which the most abundant peptides are selected for fragmentation, followed by the analysis of these fragments that will contribute to the identification of the analyzed peptides. Although a unique method of analysis has been developed for each technique, they both rely on the acquisition of mass-to-charge ratios (m/z). By the collection of this type of data, different compounds

present in a given sample can be measured on different mass spectrometry platforms that have different configurations for data ionization, separation, and detection.

The identification of proteins by mass spectrometry generally relies on genetic sequences of interest, which are translated into peptide sequences, generating a virtual spectral library. Mass spectra data (experimental spectra) are acquired, and a positive correlation between the theoretical and experimental spectra suggests the presence of peptides in a biological sample (Eng *et al.*, 1994). Due to advances in the development of new sequencing techniques, the vast amount of genetic information currently available has become attractive for species with a fully sequenced genome, as well as those where only part of this information is available. For instance, *Eucalyptus grandis* was sequenced and made publicly available, which opened numerous biotechnological perspectives (Myburg *et al.*, 2014).

Spectral correlation approaches, despite being very precise, are techniques that allow the identification of peptides only by gene sequences presented in the reference genome. As a result, it limits the identification of novel peptides and isoforms that are encoded by regions considered "non-coding" and the identification of new structures present in different open reading frames of the genome. However, it is known that the genome of eukaryotic organisms can be much more complex as it can be modified and influenced by several cellular molecular events (Barrett *et al.*, 2012).

2.5. Proteogenomics

The term proteogenomics was introduced in 2004 (Jaffe *et al.*, 2004) and usually refers to a search strategy for peptide identities derived from spectrometric data that uses customized database sequences and their construction depends on the goal of each research (Nesvizhskii, 2014). This area of study is at the interface of proteomics and genomics, and the term may include any type of application where a proteogenomics-like approach is used to interpret MS/MS spectra, which not only provide protein-level validation of gene expression and gene model refinement but also allows the enhancement of protein sequence databases (Nesvizhskii, 2014). Those novel peptides may also contain substitutions and can represent novel protein-coding loci (Nesvizhskii, 2014).

In addition, the various large-scale transcript sequencing projects are also a rich source of information about nucleotides that can be used in proteogenomic studies for different species. In plants, Baerenfaller *et al.* (2008), using a proteogenomic strategy, identified almost 50% of all proteins predicted by the *Arabidopsis thaliana* genome and identified 57 new genetic models from the identification of 261 new peptides.

Novel coding sequences may include intronic sequences, intergenic regions, pseudogenes, and alternate 5' and 3' end sequences. Castellana *et al.* (2008) observed that of the 144,079 peptides identified in an extensive analysis of spectrometric data against proteogenomic libraries, 126,055 were identified as originating from gene coding regions. Furthermore, these authors identified 18,024 new peptides, corresponding to 778 new coding regions, suggesting that the information on genetic models available for the species at the time corresponded to only 82% of the *A. thaliana* proteome.

Following a spectral correlation approach against modified RNA databases, some novel peptide genomic classifications may be identified in regions previously considered as non-coding or mapped to a protein-coding gene (Figure 1). Genomic mapping of novel peptides to protein-doing gene regions can be classified as 5’UTR, 5’UTR-EXON, Exon region in an alternative reading frame, EXON-3’UTR, 3’UTR, and dubious mappings (Figure 1).

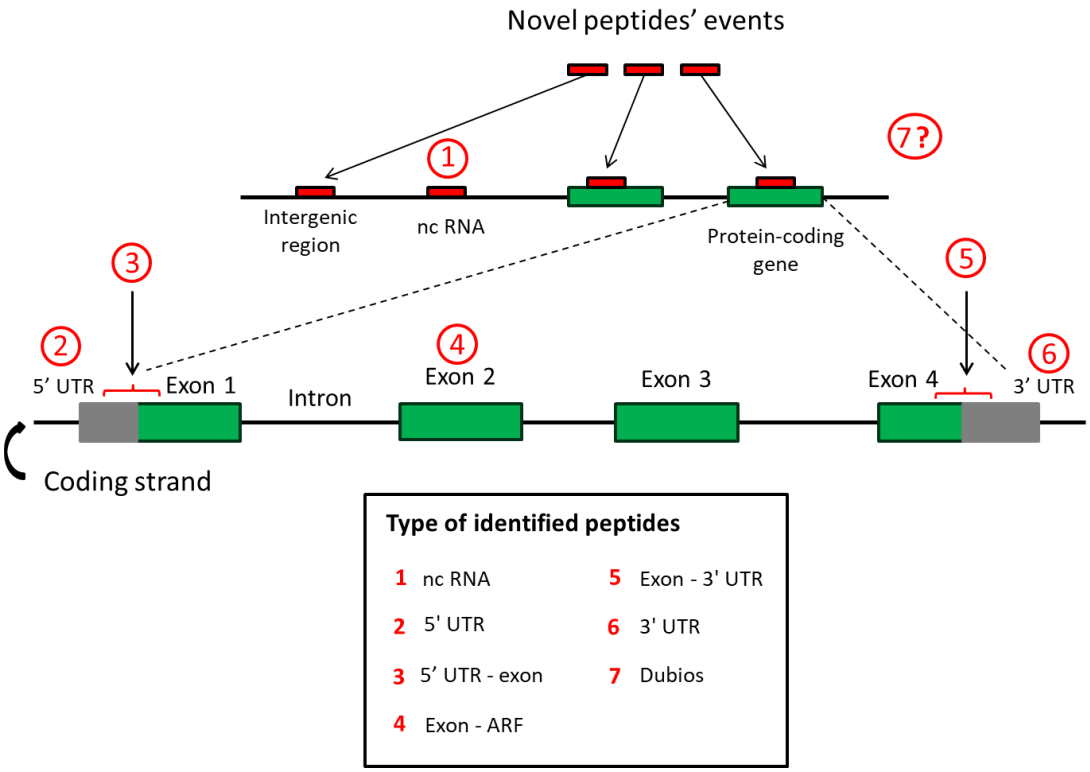


Figure 1 – Genomic classification of novel peptides identified by spectral correlations against modified RNA databases (Figure adapted from Nesvizhskii (2014)).

Some novel peptide sequences can also be identified through *de novo* peptide sequencing. *De novo* peptide sequencing is a method for peptide sequencing performed without prior knowledge of the amino acid sequence. This method can identify peptide sequences with no protein database comparison, which can overcome the limitations of database-dependent methods and deduce the sequence of peptides directly from the experimental MS/MS spectra (Standing, 2003).

Out of those *de novo* identifications, single amino acid substitutions (SAS) can also be identified through a dedicated pipeline. Automated *de novo* sequencing from MS/MS spectra in PEAKS Studio software (Bioinformatics Solutions, Waterloo, Ontario, Canada) can be used to provide *de novo* peptides and protein sequences. Subsequently, PEAKS DB, which is a database search module in PEAKS Studio, is the second step to identify Peptide-spectrum matches (PSMs) from existing protein databases (Zhang *et al.*, 2012). Lastly, aiming to detect single amino acid substitutions, datasets are searched using the SPIDER algorithm (Han et al., 2005), which is integrated within the PEAKS software.

2.6. Single amino acid substitution impact on protein stability

Single amino acid substitutions can have a high impact on protein thermodynamic stability, which can also result in protein structural changes. Those changes may be associated with mutation-induced protein biophysics, genomic variant interpretation, as well as disease-related conditions. In plants and other organisms, single amino acid substitutions may alter hydrogen bonds, salt bridges, ion-pairs, disulfide bonds, and van der Walls interactions, which may lead to changes in stability (Gromiha and Selvaraj, 2004; Kucukkal *et al.*, 2015). If stability is affected, functional changes are more likely to be predicted; however, some function-changing mutations do not affect stability (Bromberg and Rost, 2009).

Protein thermodynamic changes from single amino acid substitutions are associated with the biophysical process of protein folding between two states: before and after single amino acid substitution on a given protein. Such processes can be quantified based on the Gibbs free energy of a protein between the folding and unfolding status upon a single amino acid substitution, usually represented as ΔΔG (Li *et al.*, 2020a) As biological experiments to assess the effect of single amino acid on protein stability are usually expensive and labor extensive, computational methods are useful as they can speed up initially screening processes and provide important insights into assessing the impact of single amino acid substitutions on protein stability.

Machine learning approaches have been designed for assessing thermodynamic stability changes upon substitutions and the resulting phenomenal thermodynamic effect on proteins. For instance, some available tools have been categorized into structure-based approaches (Pires *et al.*, 2014; Savojardo *et al.*, 2016; Chen *et al.*, 2020b), sequence-based approaches (Cheng *et al.*, 2006; Folkman *et al.*, 2016; Quan *et al.*, 2016), as well as integrated approaches (Chen *et al.*, 2013; Chen *et al.*, 2020a).

Structure-based methods such as PremPS and mCSM can provide good performance for assessing the impact of single amino acid substitutions on protein stability. PremPS calculations are based on using of random forest (RT) regression scoring function and training on experimental data of unfolding Gibbs free energy changes for experimentally tested substitutions in proteins (Chen *et al.*, 2020b);

whereas, mCSM calculations are constructed upon predictive machine learning methods, which implement graph-based atom distance patterns to represent protein residue environments (Pires *et al.*, 2014).

Since not all proteins have an adequate protein structure template to be compared, sequence-based approaches can provide a good alternative for investigating the impact of single amino acid substitutions on protein stability. As a result, many high-performance methods use sequence features and information to predict protein stability changes, and a great example is MUpro (Cheng *et al.*, 2006). MUpro can predict very accurately both value and sign of energy change using support vector machine (SVM) and sequence information. This tool can be applied to many cases in which the tertiary structure of the protein is unknown, which helps overcome a major limitation for methods that require tertiary information as an input.

Modifications in the protein non-covalent interactions such as hydrophobic interactions, hydrogen bonds, Ionic interactions, and Van der Waals interactions may also affect protein stability. Such interactions can be located in the interior of a structured protein, which is generally a densely packed core of hydrophobic amino acid side chains (creating hydrophobic interactions) (Nelson and Cox, 2017). Ionic interactions also limit structural flexibility and give uniqueness to a specific protein structure that nonspecific hydrophobic interactions cannot provide (Nelson and Cox, 2017). Moreover, hydrogen bonds often have a central role in leading the protein-folding process (Nelson and Cox, 2017). Lastly, Van der Waals interactions are dipole-dipole interactions that involve the permanent electric dipoles in certain groups and are generally weak and individually contribute little to overall protein stability (Nelson and Cox, 2017).

2.7. Targeted proteomics experiments

Targeted proteomic analysis is a technique that has high reproducibility, selectivity, sensitivity, and wide dynamic range, and has been used for more accurate identification and quantification of proteins of interest, especially those present in low abundance (Gibbons *et al.*, 2019). This setup also provides high analytical reproducibility, a better signal-to-noise ratio, and increased dynamic range (Lange *et*

al., 2008). For the development of targeted proteomics methods, the first step is the determination of the group of peptides or proteins of interest for the research. The choice of the target group can be based on shotgun experiments performed or information that can be data from literature and databases (Lange *et al.*, 2008).

After determining the targets to be monitored, several targeted proteomics approaches can be applied. According to Elschenbroich and Kislinger (2011), SRM (Selected Reaction Monitoring) or MRM (Multiple Reaction Monitoring) methods use a triple quadrupole mass spectrometer. In this approach, the first quadrupole (Q1) works as a mass filter, selecting the precursor ion, which is the peptide to be monitored. The second quadrupole (Q2) is a collision cell, in which the selected peptide will undergo fragmentation by the CID (Collision-induced dissociation) method. Lastly, the third quadrupole (Q3), performs the same function as Q1, however, it selects the fragment ion that will be directed to the detector.

For the SRM method, it is necessary to define both the peptide and the fragment ions to be analyzed. The m/z pair of the precursor ion and the fragment ion is called a transition. The choice of transitions should be based mainly on the higher signal generated by the fragment ion with the least possible interference since the complexity of biological samples at the protein level may hinder the detection of specific transitions (Janecki *et al.*, 2007). These transitions must be validated to confirm the peak identity. Another possibility is the use of synthetic peptides for the optimization and validation of transitions (Picotti and Aebersold, 2012). Multiplexed MRM-based applications were used for biomarker discoveries, validation, and accurate quantification of human disease studies (Freue and Borchers, 2012; Percy *et al.*, 2013), as well as to investigate protein biomarkers in Loquat fruits for polyphenol oxidase (Martínez-Márquez *et al.*, 2013).

The PRM (Parallel Reaction Monitoring) is another method that can be applied using high-resolution mass spectrometers as hybrid quadrupole-Orbitrap and quadrupole-time-of-flight instruments. In these instruments, the peptide to be analyzed is selected by the quadrupole, which works as the mass filter, and this ion is directed to a collision cell where it will undergo fragmentation, and lastly, all fragments will be analyzed in Orbitrap. In this approach, it is then possible to obtain the MS2 spectrum of the target peptide, without previously defining the transitions to be analyzed, as in

the case of SRM (Ronsein *et al.*, 2015). In general, PRM and SRM exhibited comparable linearity, dynamic range, precision, and repeatability for protein quantification (Ronsein *et al.*, 2015).

3. REFERENCES

Baerenfaller K, Grossmann J, Grobei MA, Hull R, Hirsch-Hoffmann M, Yalovsky S, Zimmermann P, Grossniklaus U, Gruissem W, Baginsky S (2008) Genome-scale proteomics reveals *Arabidopsis thaliana* gene models and proteome dynamics. **Science** 320: 938-941.

Barrett LW, Fletcher S, Wilton SD (2012) Regulation of eukaryotic gene expression by the untranslated gene regions and other non-coding elements. **Cellular and Molecular Life Sciences** 69: 3613–3634.

Bromberg Y, Rost B (2009) Correlating protein function and stability through the analysis of single amino acid substitutions. **BMC Bioinformatics** 10: (Suppl 8). doi: 10.1186/1471-2105-10-S8-S8

Castellana NE, Payne SH, Shen Z, Stanke M, Bafna V, Briggs SP (2008) Discovery and revision of *Arabidopsis* genes by proteogenomics. **Proceedings of the National Academy of Sciences** 105 (52): 21034-21038.

Castro CA de O, Resende RT, Bhering LL, Cruz CD (2016) Breve histórico do melhoramento genético do eucalipto no Brasil sob a ótica dos avanços biométricos. **Ciencia Rural** 46: 1585–1593.

Chai Q, Gan Y, Zhao C, Xu HL, Waskom RM, Niu Y, Siddique KHM (2016) Regulated deficit irrigation for crop production under drought stress. A review. **Agronomy for Sustainable Development** 36: 1–21.

Chen CW, Lin J, Chu YW (2013) iStable: Off-the-shelf predictor integration for predicting protein stability changes. **BMC Bioinformatics** 4 (Suppl 2): S5. doi: 10.1186/1471-2105-14-S2-S5

Chen CW, Lin MH, Liao CC, Chang HP, Chu YW (2020a) iStable 2.0: Predicting protein thermal stability changes by integrating various characteristic modules. **Computational and Structural Biotechnology Journal** 18: 622–630.

Chen Y, Lu H, Zhang N, Zhu Z, Wang S, Li M (2020b) PremPS: Predicting the impact of missense mutations on protein stability. **PLoS Computational Biology** 16(12): e1008543. doi: 10.1371/journal.pcbi.1008543

Cheng J, Randall A, Baldi P (2006) Prediction of protein stability changes for single-site mutations using support vector machines. **Proteins: Structure, Function and Genetics** 62: 1125–1132.

Easterling DR, Meehl GA, Parmesan C, Changnon SA, Karl TR, Mearns LO (2000) Climate Extremes: Observations, Modeling, and Impacts. **Science** 289: 2068-2074.

Elschenbroich S, Kislinger T (2011) Targeted proteomics by selected reaction monitoring mass spectrometry: Applications to systems biology and biomarker discovery. **Molecular BioSystems** 7: 292–303.

Eng JK, Mccormack AL, Yates JR (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. **Journal of the American Society for Mass Spectrometry** 11: 976-89.

Feist P, Hummon AB (2015a) Proteomic challenges: Sample preparation techniques for Microgram-Quantity protein analysis from biological samples. **International Journal of Molecular Sciences** 16: 3537–3563.

Feist P, Hummon AB (2015b) Proteomic challenges: Sample preparation techniques for Microgram-Quantity protein analysis from biological samples. **International Journal of Molecular Sciences** 16: 3537–3563.

Folkman L, Stantic B, Sattar A, Zhou Y (2016) EASE-MM: Sequence-Based Prediction of Mutation-Induced Stability Changes with Feature-Based Multiple Models. **Journal of Molecular Biology** 428: 1394–1405.

Freue GVC, Borchers CH (2012) Multiple Reaction Monitoring (MRM): Principles and application to coronary artery disease. **Circulation: Cardiovascular Genetics** 3: 378. doi: 10.1161/CIRCGENETICS.111.959528

Gao X, Li H, Li H, Dong S, Chu J, Guo H, Zhao Q (2018) Sensitive determination of nine anticoagulant rodenticides in blood by high resolution mass spectrometry with supported liquid extraction pretreatment. **Forensic Science International** 292: 39–44.

Gibbons BC, Fillmore TL, Gao Y, Moore RJ, Liu T, Nakayasu ES, Metz TO, Payne SH (2019) Rapidly Assessing the Quality of Targeted Proteomics Experiments through Monitoring Stable-Isotope Labeled Standards. **Journal of Proteome Research** 18: 694–699.

Gromiha MM, Selvaraj S (2004) Inter-residue interactions in protein folding and stability. **Progress in Biophysics and Molecular Biology** 86: 235–277.

Han Y, Ma B, Zhang K (2005) SPIDER: Software for Protein Identification from Sequence Tags with De Novo Sequencing Error. **Journal of Bioinformatics and Computational Biology** 3 :697-716.

IBÁ - Indústria Brasileira de Árvores (2020) **Annual Report**. Brasília, 66 p.

Jaffe JD, Berg HC, Church GM (2004) Proteogenomic mapping as a complementary method to perform genome annotation. **Proteomics** 4: 59–77.

Janecki DJ, Bemis KG, Tegeler TJ, Sanghani PC, Zhai L, Hurley TD, Bosron WF, Wang M (2007) A multiple reaction monitoring method for absolute quantification of the human liver alcohol dehydrogenase ADH1C1 isoenzyme. **Analytical Biochemistry** 369: 18–26.

Jiménez S, Dridi J, Gutiérrez D, Moret D, Irigoyen JJ, Moreno MA, Gogorcena Y (2013) Physiological, biochemical and molecular responses in four *Prunus* rootstocks submitted to drought stress. **Tree Physiology** 33: 1061–1075.

Kucukkal TG, Petukh M, Li L, Alexov E (2015) Structural and physico-chemical effects of disease and non-disease nsSNPs on proteins. **Current Opinion in Structural Biology** 32: 18–24.

Lange V, Picotti P, Domon B, Aebersold R (2008) Selected reaction monitoring for quantitative proteomics: A tutorial. **Molecular Systems Biology** 4: 222. doi: 10.1038/msb.2008.61

Li B, Yang YT, Capra JA, Gerstein MB (2020a) Predicting changes in protein thermodynamic stability upon point mutation with deep 3D convolutional neural networks. **PLoS Computational Biology** 16(11): e1008291. doi: 10.1371/journal.pcbi.1008291

Li H, Li Y, Ke Q, Kwak SS, Zhang S, Deng X (2020b) Physiological and differential proteomic analyses of imitation drought stress response in sorghum bicolor root at the seedling stage. **International Journal of Molecular Sciences** 21: 1–28.

Marchiori JNC (2014) Primórdios da Silvicultura no Rio Grande do Sul. 1- **NOTA SOBRE A INTRODUÇÃO DO GÊNERO EUCALYPTUS L'HER. BALDUINA** 30–III: 21–31.

Martínez-Márquez A, Morante-Carriel J, Sellés-Marchart S, Martínez-Esteso MJ, Pineda-Lucas JL, Luque I, Bru-Martínez R (2013) Development and validation of MRM methods to quantify protein isoforms of polyphenol oxidase in loquat fruits. **Journal of Proteome Research** 12: 5709–5722.

Martins GS, Freitas NC, Máximo WPF, Paiva LV (2018) Gene expression in two contrasting hybrid clones of *Eucalyptus camaldulensis x Eucalyptus urophylla* grown under water deficit conditions. **Journal of Plant Physiology** 229: 122–131.

Myburg AA, Grattapaglia D, Tuskan GA, Hellsten U, Hayes RD, Grimwood J, Jenkins J, Lindquist E, Tice H, Bauer D, et al (2014) The genome of *Eucalyptus grandis*. **Nature** 510: 356–362.

Nelson DL, Cox MM (2017) **Princípios de Bioquímica de Lehninger**, 7th ed. W.H. Freeman, 1312 p.

Nesvizhskii AI (2014) Proteogenomics: Concepts, applications and computational strategies. **Nature Methods** 11: 1114–1125.

Ngara R, Ramulifho E, Movahedi M, Shargie NG, Brown AP, Chivasa S (2018) Identifying differentially expressed proteins in sorghum cell cultures exposed to osmotic stress. **Scientific Reports** 8: 8671. doi: 10.1038/s41598-018-27003-1

Percy AJ, Chambers AG, Yang J, Borchers CH (2013) Multiplexed MRM-based quantitation of candidate cancer biomarker proteins in undepleted and non-enriched human plasma. **Proteomics** 13: 2202–2215.

Picotti P, Aebersold R (2012) Selected reaction monitoring-based proteomics: Workflows, potential, pitfalls and future directions. **Nature Methods** 9: 555–566.

Pires DEV, Ascher DB, Blundell TL (2014) MCSM: Predicting the effects of mutations in proteins using graph-based signatures. **Bioinformatics** 30: 335–342.

Quan L, Lv Q, Zhang Y (2016) STRUM: Structure-based prediction of protein stability changes upon single-point mutation. **Bioinformatics** 32: 2936–2946.

Ronsein GE, Pamir N, von Haller PD, Kim DS, Oda MN, Jarvik GP, Vaisar T, Heinecke JW (2015) Parallel reaction monitoring (PRM) and selected reaction monitoring (SRM) exhibit comparable linearity, dynamic range and precision for targeted quantitative HDL proteomics. **Journal of Proteomics** 113: 388–399.

SAMPAIO AN (1961) História. In: ANDRADE, E.N. **O Eucalipto**. Cia Paulista de Esgradas de Ferro, Sao Paulo, 665 p.

Savojardo C, Fariselli P, Martelli PL, Casadio R (2016) INPS-MD: A web server to predict stability of protein variants from sequence and structure. **Bioinformatics** 32: 2542–2544.

Shao HB, Chu LY, Jaleel CA, Manivannan P, Panneerselvam R, Shao MA (2009) Understanding water deficit stress-induced changes in the basic metabolism of higher plants-biotechnologically and sustainably improving agriculture and the ecoenvironment in arid regions of the globe. **Critical Reviews in Biotechnology** 29: 131–151.

Shi T, Song E, Nie S, Rodland KD, Liu T, Qian WJ, Smith RD (2016) Advances in targeted proteomics and applications to biomedical research. **Proteomics** 16: 2160–2182.

Standing KG (2003) Peptide and protein de novo sequencing by mass spectrometry. **Current Opinion in Structural Biology** 13: 595–601.

Staudinger MD, Carter SL, Cross MS, Dubois NS, Duffy JE, Enquist C, Griffis R, Hellmann JJ, Lawler JJ, O'Leary J, et al (2013) Biodiversity in a changing climate: A synthesis of current and projected trends in the US. **Frontiers in Ecology and the Environment** 11: 465–473.

Teng S, Srivastava AK, Schwartz CE, Alexov E, Wang L (2010) Structural assessment of the effects of Amino Acid Substitutions on protein stability and protein-protein interaction. **International Journal of Computational Biology and Drug Design** 3: 334–349.

Zhang J, Xin L, Shan B, Chen W, Xie M, Yuen D, Zhang W, Zhang Z, Lajoie GA, Ma B (2012) PEAKS DB: De novo sequencing assisted database search for sensitive and accurate peptide identification. **Molecular and Cellular Proteomics** 4: M111.010587. doi: 10.1074/mcp.M111.010587

Zhou S, Medlyn B, Sabaté S, Sperlich D, Colin Prentice I (2014) Short-term water stress impacts on stomatal, mesophyll and biochemical limitations to photosynthesis differ consistently among tree species from contrasting climates. **Tree Physiology** 34: 1035–1046.

CHAPTER 2 - Identification of novel protein-coding sequences in *Eucalyptus grandis* plants by high-resolution mass spectrometry[1]

[1]This chapter corresponds to the scientific article published in the Journal Biochimica et Biophysica Acta - Proteins and Proteomics 1869(3):140594, 2021 (doi: 10.1016/j.bbapap.2020.140594).

Gabriel Lemes Jorge[1], Tiago Santana Balbuena[1]*

[1] Sao Paulo State University, Department of Technology, Jaboticabal, Sao Paulo, Brazil

*Corresponding author: Tiago Santana Balbuena
Contact: Via de Acesso Prof. Paulo Donato Castellane. Universidade Estadual Paulista “Júlio de Mesquita Filho”, UNESP, Departamento de Tecnologia, Jaboticabal, SP, Brazil, 14884-900.

Gabriel Lemes Jorge - ORCID: https://orcid.org/0000-0003-3541-5129
Email: gabriel.lemes@unesp.br
Tiago Santana Balbuena - ORCID: https://orcid.org/0000-0002-1053-0254
Email: tiago.balbuena@unesp.

ABSTRACT

Eucalyptus species are widely used in the forestry industry, and a significant increase in the number of sequences available in database repositories has been observed for these species. In proteomics, a protein is identified by correlating the theoretical fragmentation spectrum derived from genomic/transcriptomic data against the experimental fragmentation mass spectrum acquired from large-scale analysis of protein mixtures. Proteogenomics is an alternative approach that can identify novel proteins encoded by regions previously considered as non-coding. This study aimed to confidently identify and confirm the existence of previously unknown protein-coding sequences in the *Eucalyptus grandis* genome. To this end, we used a modified spectral correlation strategy and a dedicated *de novo* peptide sequencing pipeline. Upon the strategy used here, we confidently identified 41 novel peptide forms and six peptides containing at least one single amino acid substitution. The most representative genomic class of novel peptides was identified as originating from alternative reading frames. In contrast, no clear single amino acid substitution pattern was identified. Validation of the identifications was carried out using a parallel reaction monitoring approach that provided further mass spectrometry support for the existence of the novel peptide sequences. Data are available via ProteomeXchange with identifier PXD022110.

Keywords: bottom-up proteomics, parallel reaction monitoring, proteogenomics, proteomics.

1. INTRODUCTION

Eucalyptus species are widely used in the forestry industry because of their rapid growth and wood properties for cellulose and paper production. Due to advances in the development of new sequencing techniques, the vast amount of gene information currently available has become attractive for species with a fully sequenced genome and for those in which only a part of this information is available. The genome of *Eucalyptus grandis* has been sequenced and made publicly available (Myburg *et al.* 2014), which opened numerous biotechnological perspectives and accelerated proteomics studies involving this species.

In contrast to sequencing technologies, protein identification is primarily performed as a result of correlating mass spectra acquired by processing a biological sample (experimental spectra) with those virtually generated by processing sequences encoded in a reference genome (virtual spectra). The positive correlation observed between the theoretical and the experimentally obtained spectra is strong evidence of the presence of a particular peptide in biological samples (Eng *et al.* 1994). Additionally, another strategy for protein identification is to perform *de novo* peptide sequencing, which directly infers the peptide sequence from an acquired tandem mass spectrum with no comparison with a reference database (Zhang *et al.* 2011).

Currently, it is known that the genomic expression in eukaryotic organisms is much more dynamic than previously reported. In the conventional spectral correlation approach, the protein-coding sequences are known, correctly inferred, and present in this reference genomic library. However, such restriction limits the identification of new proteins and isoforms encoded by regions considered as "non-coding", and the identification of novel peptide forms of that genome are consequently compromised. As a result, a powerful method known as proteogenomics has emerged, and it may assist with identifying novel coding regions and peptide forms.

The term proteogenomics was first introduced in the literature in 2004 (Jaffe *et al.* 2004). It usually refers to the strategy of searching for peptide identities derived from spectrometric data using custom databases, where their construction depends on the objective of the ongoing research project. For that to occur, the genomic information must be processed in its different possibilities and integrated with other

gene product information such as transcripts. In addition, the various large-scale transcript sequencing projects are also rich sources of nucleotide information that can be used in proteogenomics studies as virtual spectra for spectral correlations. Another approach that could enhance the identification of novel peptide forms is *de novo* peptide sequencing, since peptides may contain single amino acid substitutions originated from single nucleotide polymorphisms (SNPs) (Nesvizhskii 2014). In recent years, *de novo* peptide sequencing has gained popularity and proved to be a valuable strategy for deep-proteome profiling experiments and functional analysis from non-model organisms (Blank-Landeshammer *et al.* 2017).

The combination of proteogenomics and *de novo* peptide sequencing analysis can provide a more comprehensive view of the proteome in eukaryotic organisms, as well as improve genome annotation quality (Blank-Landeshammer *et al.* 2019). Proteogenomics has many applications, such as in clinical and cancer-related studies (Ang *et al.* 2019), genome annotation improvement (Potgieter *et al.* 2016) and in enhancing the understanding of phenotypic variation (Heunis *et al.* 2017). In plants, proteogenomics has contributed to identifying novel peptides and coding regions in the *Arabidopsis thaliana* genome (Castellana *et al.* 2008; Baerenfaller *et al.* 2008), as well as to improving genome annotation in grapevine (Chapman and Bellgard 2017).

Confidently identifying novel coding events in an organism is important, but validating such findings is also greatly useful. As an alternative strategy to validate the identification of novel peptides and single amino acid substitutions, the parallel reaction monitoring (PRM) approach is commonly used. PRM-based targeted methods provide high specificity, as the tandem mass spectrometry (MS/MS) data are acquired in high-resolution mode and high mass accuracy that can distinguish background ions from target peptide ions (Peterson *et al.* 2012; Rauniyar, 2015). With its high resolution, the PRM method can identify compounds with greater sensitivity and selectivity (Gao *et al.* 2018). This method has also been used in other studies to validate protein biomarkers (Sandow *et al.* 2018) and identify novel peptides in prokaryotic organisms (Omasits *et al.* 2017).

This presented strategy can identify novel peptide forms that cannot be identified using conventional protein identification workflows, which usually are based on spectral correlations against a reference protein database. Therefore, this study

aimed to identify previously unknown protein-coding sequences in the *E. grandis* genome by identifying and validating novel peptide sequences using a dedicated proteogenomics approach.

2. MATERIAL AND METHODS

2.1. Plant growth and sample preparation

Eucalyptus grandis seeds were acquired from the Institute of Research and Forestry Studies (Instituto de Pesquisa e Estudos Florestais - IPEF) (Cultivar: LCFA001; Seed lot: BO0027N01). Plant germination was carried out in 1 L pots containing a commercial substrate (Bioplant®). After four weeks of growth, the seedlings were transplanted to 6 L pots and the young *E. grandis* plants were cultivated in greenhouse at the Department of Technology at the São Paulo State University (Jaboticabal - SP, Brazil). During cultivation, cultural practices such as fertilization, phytosanitary control, and irrigation were carried out according to the plants' need. When plants were 90 days old, middle stem portions of nine plants were collected, immediately frozen in liquid nitrogen, and stored at -80 °C prior to protein extraction.

2.2. Protein Extraction and Quantification

Two grams of each plant stem material were individually processed, macerated with liquid nitrogen using a pestle until a fine powder, and transferred to 15 mL clean tubes. Each sample was extracted through addition of 12 mL Tris-HCl buffer solution (125 mM Tris pH 6.8, 20% glycerol, 1% SDS, 1% DTT). After centrifugation at 3200 xg for 20 min (4 °C), the supernatant was recovered and protein isolation was performed through overnight precipitation (-30 °C) upon addition of an organic solvent. The protein pellet was then resuspended in 150 µL buffer solution and the soluble protein concentration was determined using the Bio-Rad Protein Assay kit according to the manufacturer's instructions.

2.3. Polyacrylamide gel electrophoresis and protein digestion

Prior to mass spectrometry analysis, a 50 µg aliquot of stem proteins was loaded into sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) gel and shortly separated for 1 h. The resulting 2 cm-wide band was then sliced into small gel pieces and transferred to clean tubes to be reduced, alkylated and, finally, *in-gel* digested (Shevchenko *et al.* 2007). Protein digestion was performed by adding a volume of trypsin aiming at an enzyme/protein ratio of 1:10 at 37 °C for 16 h. After digestion, the tryptic peptides were eluted from the gel matrix and dried under vacuum centrifugation. Subsequently, the digested peptides were submitted to two filtering steps: the first filtering was carried out using Costar® Spin-X® centrifuge tube filters (Salt Lake City, Utah, USA) and the second sample clean-up was performed using a Millipore C18 Ziptip® (Billerica, MA, USA), according to the manufacturers' instructions.

2.4. Acquisition of mass spectrometric data

After tryptic digestion, all nine samples were combined into one pool sample. The peptide mixture was resuspended in 180 µL (0.1% formic acid) and 10 µL was introduced in a single injection into an Easy nLC 1000, Thermo Scientific® system. The peptides were separated by reverse-phase liquid chromatography using a nanoLC column (AcclaimPepMap100, C18, 3 µm, 100 Å, 75 µm id × 15 cm) and a linear gradient under a constant flow of 400 nL min^{-1}. The gradient and time of chromatographic separation (95 min) were defined according to trial tests. Acquisition of high-resolution mass spectra was carried out concomitantly to the chromatographic separation. The spectra were acquired in a Q-Exactive™ (Thermo Scientific) system operated in the data-dependent acquisition (DDA) mode for the 20 most abundant peptide ions. Full (400-1600 m/z) and data-dependent scans were acquired with resolution equal to 70,000 and 17,500 FWHM (full width at half maximum), respectively. Fragment spectra were acquired upon HCD (high-energy collision dissociation) fragmentation of the isolated peptide ions under normalized collision energy of 30 eV. The duty cycle time was optimized through changes in the automatic

gain control parameters of the instrument. The peptide ions selected for fragmentation were excluded in a new isolation event for 5 s.

2.5. Identification of novel peptides by spectral correlations

- Modified database creation

Sequences from the following *E. grandis* National Center for Biotechnology Information (NCBI) RefSeq databases (GCF_000612305.1) were downloaded and virtually translated into the three possible reading frames using the publicly available Virtual Ribosome v.2.0 (Wernersson R. 2006) software: mRNA (53,109 entries), ncRNA (8,388 entries), and transcribed RNA (4,172 entries). In order to minimize the false identifications resulting from the different database sizes used here, the *E. grandis* protein database from NCBI: txid71139 (99,946 entries) was combined with each modified database, generating nine chimeric databases that were used for identification purposes.

- Novel peptide identification

Novel peptide identification was performed using the spectral correlation approach. The candidate sequences were identified using a dedicated pipeline with stringent controls of false-discovery rates. Peptide identification was carried out using the Comet searching tool embedded in the PatternLab for Proteomics platform (Carvalho *et al.* 2008). Aiming to define a confidence threshold for peptide spectrum matches (PSMs), a target-decoy approach was used to estimate the false discovery rate (FDR). Cysteine carbamidomethylation (Cys) and methionine oxidation (Met) were set as static and dynamic amino acid mass modifications, respectively. Additionally, precursor and fragment mass tolerance values of 20 ppm and 0.02 Da, respectively, and two missed cleavages per peptide were allowed for confident peptide identification. All spectral peptide alignments were filtered using the SePro software (CARVALHO *et al.* 2012, 2016) and adjusted for positive FDR with a limit of 1% for peptide identification. An additional proteome mining was performed by manually

checking whether novel peptides were previously present in the *E. grandis* v.2.0 protein database (46,280 entries) before inclusion in the final list.

- Novel peptide mapping and genomic classification

Aiming to identify orthologous peptide sequences, a protein–protein basic local alignment search tool (*blastp*) was used in each of the high-scoring novel peptides against all sequences in the GenBank non-redundant protein database. Additionally, *tblastn* searches against the *E. grandis* EST database were carried out to further categorize and investigate the existence of the identified novel peptides in previous cDNA libraries. For genomic classification of the novel peptides, *tblastn* searches were conducted against the *E. grandis* v.1.0 database - reference assembly in Annotation Release 101. Subsequently, genomic mapping and visualization were performed using a Web-based tool provided by NCBI, Genome Data Viewer (GDV), which is embedded in the *tblastn* result page. Searches were also performed using the default parameters of the *tblastn* search algorithm. The expected (E) values for all *tblastn* results were individually examined to identify significant blast hits (100% identity and 100% query coverage). Alignment between the novel identified peptide and the genomic location of the aligned peptide was only considered successful if both protein *loci* were the same (matching accession codes). Whenever the novel peptide was mapped to an exon region, this genomic mapping classification was only considered if the novel peptide reading frame identification was different from the longest reading frame for the respective NCBI protein sequence.

Taxonomic assignment and functional annotation of novel peptides were performed following the eggNOG (Evolutionary genealogy of genes: Non-supervised Orthologous Groups) databases (v5) pipeline (Huerta-Cepas *et al.* 2019). The predicted protein-coding sequence used in the search was derived from each NCBI RefSeq RNA accession code. For novel peptides identified in RefSeq ncRNA and transcribed RNA databases (no NCBI predicted protein), the protein sequence from the in-silico translation frame was used instead. Searches were performed with no orthologous group restrictions. The highest-scoring annotated hit was selected for each novel peptide functional annotation.

2.6. Identification of single amino acid substitution by *de novo* sequencing

The PEAKS Studio 10.5 (Bioinformatics Solutions, Waterloo, Ontario, Canada) software was used to provide automated *de novo* sequencing from MS/MS spectra. Data refinement of the *E. grandis* stem raw file was carried out prior to *de novo* sequencing by correcting precursor scans according to mass and charge states (minimum and maximum charge states = 2 and 4, respectively). PEAKS *de novo* sequencing was performed with precursor and fragment error tolerance values of 15 ppm and 0.01 Da, respectively. Trypsin was used as protease with zero maximum missed cleavage allowed. Cys and Met were set as fixed and variable modifications, respectively. A maximum of four variable modifications per peptide was allowed. Subsequently, PEAKS DB, which is a database search module in PEAKS Studio 10.5, was used in the second step to identify PSMs from existing protein databases (Zhang *et al.* 2011). The *E. grandis* v.2.0 protein database (46,280 entries) retrieved from the Phytozome website was used as reference database. To determine a confidence threshold for PSMs, the PEAKS-embedded target-decoy approach, “decoy fusion”, was used to estimate the FDR of the PEAKS DB result. Other search parameters were kept the same as previously described in the section: Identification of single amino acid substitution by de novo sequencing.

Aiming to detect single amino acid variants, the datasets were searched using the SPIDER software (Han *et al.* 2005), integrated with the PEAKS software, against the *E. grandis* reference database. The PSMs reported by SPIDER were filtered to 1%. The *de novo* only score was set to 80%, and post-translational modifications (PTMs) scores were set to ≥20. In addition, scores for all identified proteins using the PEAKS/SPIDER algorithm were given as a $-10_{\log}$ *p*-value, with a minimum score of 30 and at least a single and significant peptide per protein group. A single amino acid substitution (SAS) was only considered significant if the following criteria were met: a) minimum of two spectra containing a SAS and one valid spectrum from the PEAKS DB search for the same peptide sequence; b) spectra with well-defined “b” or “y” series at the position of the substituted amino acid; c) substitutions from asparagine to aspartic acid and from glutamine to glutamic acid were not considered, since they can result from deamination; d) mutation ion intensity of at least 10%. Therefore, all peptide

matches from both PEAKS DB and SPIDER search modules were manually verified to ensure quality before being included as valid SASs in the final peptide list.

2.7. Validation of novel peptides and single amino acid substitutions

The validation step was based on a PRM assay, which was performed by nanoflow liquid chromatography-tandem mass spectrometry (LC-MS/MS) on an Easy nLC 1000 system (Thermo Fisher Scientific) coupled to a Q-Exactive mass spectrometer (Thermo Fisher Scientific) using a Nanospray Flex™ Ion Source (Thermo Fisher Scientific).

High-energy collision dissociation (HCD) fragmentation of the target peptides was adjusted to 30 eV, after being experimentally confirmed by successive mass analyses using different collision energies. It is worth mentioning that chromatographic separation was optimized for the targeted proteomics experiment. Therefore, the discovery proteomics separation profile was initially used as a reference to optimize the chromatographic separation for all selected peptides in a single run. To this end, different gradients were evaluated, and the *m/z* signals of both precursor and fragment ions were monitored throughout the analysis. Upon optimization, acquisition of at least three *m/z* signals of fragment ions throughout the peptide chromatographic peak was expected.

Peptide mixtures were separated by reverse-phase chromatography on a gradient at a 400 nL min^{-1} constant flow rate for 95 min. The PRM method combined two scan events, starting with a full scan event followed by targeted MS/MS scans (2^+ and 3^+ charges) at a resolution of 17,500. All novel peptides and SASs selected for validation were targeted in a 4 *m/z* isolation window around the *m/z* signal of each peptide of interest. An automatic gain control (AGC) target value of $1e^5$, and a maximum injection fill time (IT) of 100 ms were set. Additionally, the retention time (RT) of each novel peptide identification was established based on the previous DDA experiment according to equation RT = x ±1.5 min, where “x” corresponds to the elution time of the target analyte.

3. RESULTS

Our proteogenomics workflow allowed identification of 41 novel peptides and six known peptides containing at least one SAS (Fig. 1). Genomic mapping of the 41 novel peptides indicated the existence of previously unknown coding sequences within the *E. grandis* genes, which allowed 44 unique and one dubious mapping assignment. Novel peptides could be identified at different rates among all modified RefSeq databases created. There were 36 novel peptides identified in the mRNA database (53,109 entries; at a 0.07 % rate), 3 novel peptides identified in the ncRNAdatabase (8,388 entries; at a 0.04 % rate), and 6 novel peptides identified in the transcribed RNA database (4,172 entries; at a 0.14 % rate).

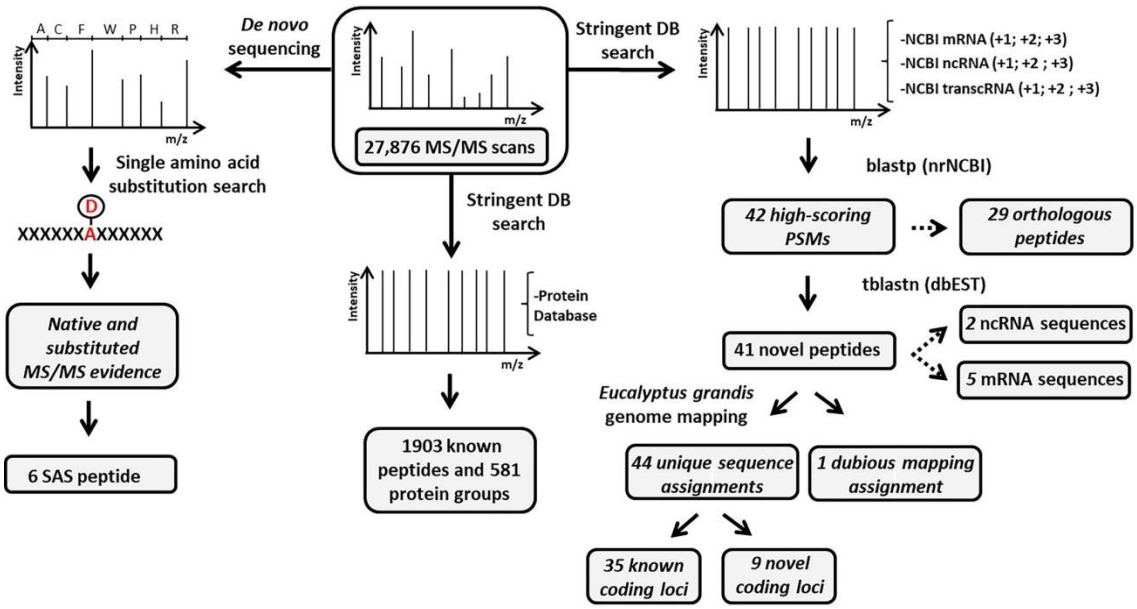


Fig. 1 – Identification of new protein-coding sequences from stems of young Eucalyptus grandis plants. Identification of single amino acid substitutions was carried out through a de novo peptide sequencing approach, whereas novel peptide identifications were performed by spectral correlations against modified RNA databases translated into three different reading frames.

Depending on the genomic location of the new coding sequences, the novel peptides were classified into the following categories (Fig. 2): 5' UTR, 5' UTR-EXON, EXON-ARF, INTRON-EXON, EXON-3' UTR, 3' UTR, NCL, and DUBIOUS.

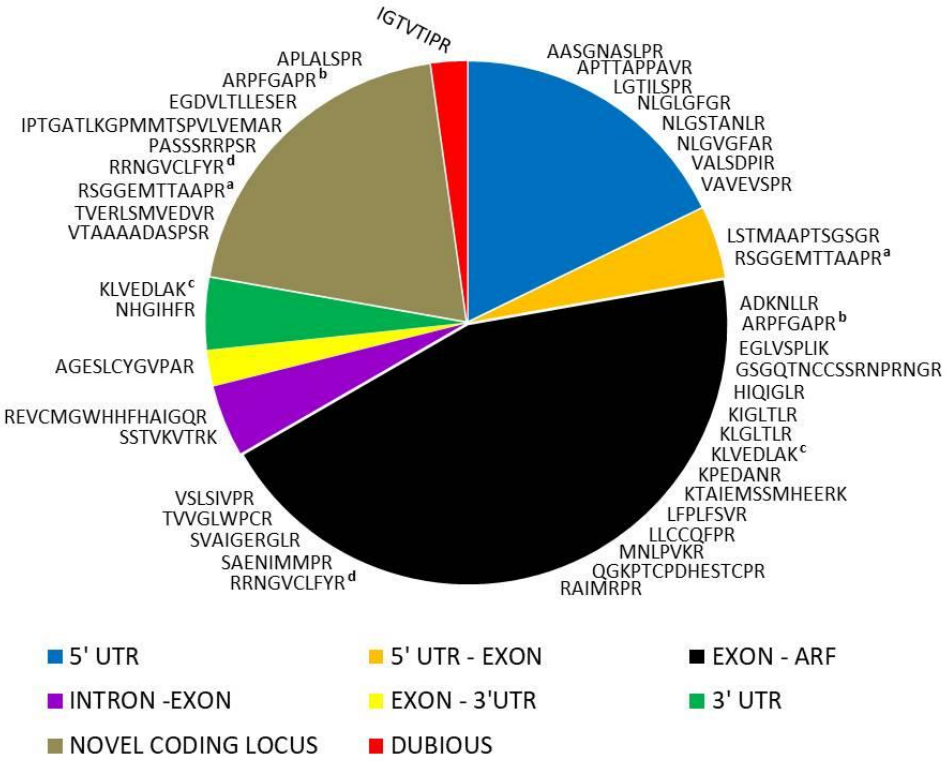


Fig. 2 – Genomic classification of novel peptides identified by spectral correlations against the modified RNA databases (mapping details can be found in Table S1). Uppercase letters indicate peptides with more than one classification assignment.

The most representative category of novel peptides was the EXON-ARF class (alternative reading frame located in an exon region) (Fig. 3), followed by the class of peptides translated from novel coding locus (NCL). The *tblastn* searches against the *E. grandis* EST database suggested additional evidence on the existence of seven RNA sequences that resulted in novel peptide identifications, from which two were translated from ncRNA sequences and five from mRNA sequences (Table S2). The functional characterization and orthologous group identification of novel peptides carried out through EggNOG pipeline revealed no clear pattern of potential functions associated with the novel peptide identifications (Table S3). However, most orthologous proteins were identified in *Prunus mume*, which is an Asian tree species.

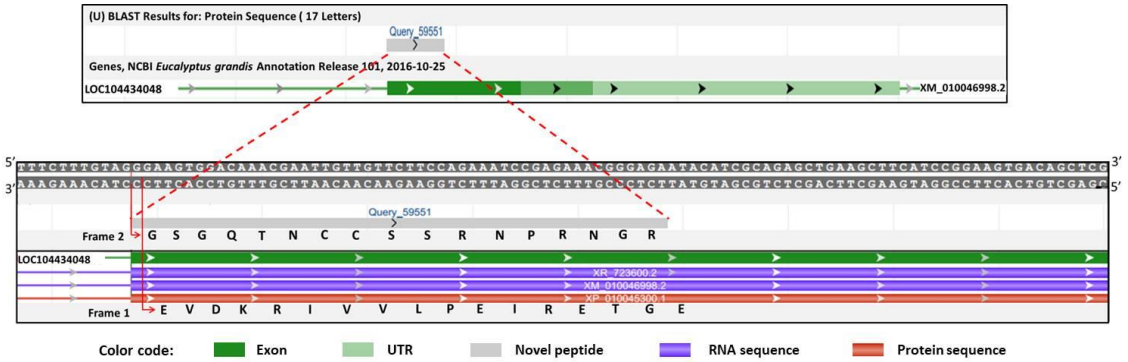


Fig. 3 – Snapshot of the genomic localization of a novel peptide translated from an alternative reading frame sequence within an exon region (EXON-ARF). Other examples of genomic localization can be found in Fig. S1.

Regarding the SASs identified through the peptide *de novo* approach, it can be inferred that no clear substitution patterns could be pinpointed from the identified single amino acid modifications (Table 1). As mentioned in the Material and Methods section, a PRM approach was used to validate the existence of the *de novo* SAS peptides. Pairwise comparisons between the peptide fragmentation patterns from the data-dependent mass spectrum and the PRM spectrum support the existence of the amino acid substituted sequences (Fig. 4).

Table 1 – Single amino acid substitutions (SAS) identified from stems of young *Eucalyptus grandis* plants through a *de novo* peptide sequencing approach.

Peptide number	Phytozome Accession Code	Sequence		Score[c]		Number of scans[d]	
		Native[a]	Substituted[b]	Native	Substituted	Native	Substituted
1	Eucgr.J03033.1	AGLQFPVGR [e]	AGLQYPVGR	62.67	28.59	19	2
2	Eucgr.J01562.1	EAPPGEKPEPVR	EAPAGEKPEPVR	79.56	45.43	6	2
3	Eucgr.B02878.1	FTQANSEVSALLGR	FVQAGSEVSALLGR	60.23	33.11	9	9
4	Eucgr.K01508.1	YGAGIGPGVYDIHSPR	YGSGIGPGVYDIHSPR	78.56	48.37	26	6
5	Eucgr.J00025.1	TTPSYVAFTDTER	TTPSYVSFTDTER	83.57	38.42	11	2
6	Eucgr.K01666.1	YIGTSELQLER	YTGTSDLQLER	29.3	21.79	2	4

[a] Peptide sequence obtained from PEAKS database search; [b] Peptide sequence obtained from PEAKS Spider Homology Search; [c] Score in -10logP scale; [d] number of MS/MS scans of each identified peptide; [e] red letters indicate the location of single amino acid substitution.

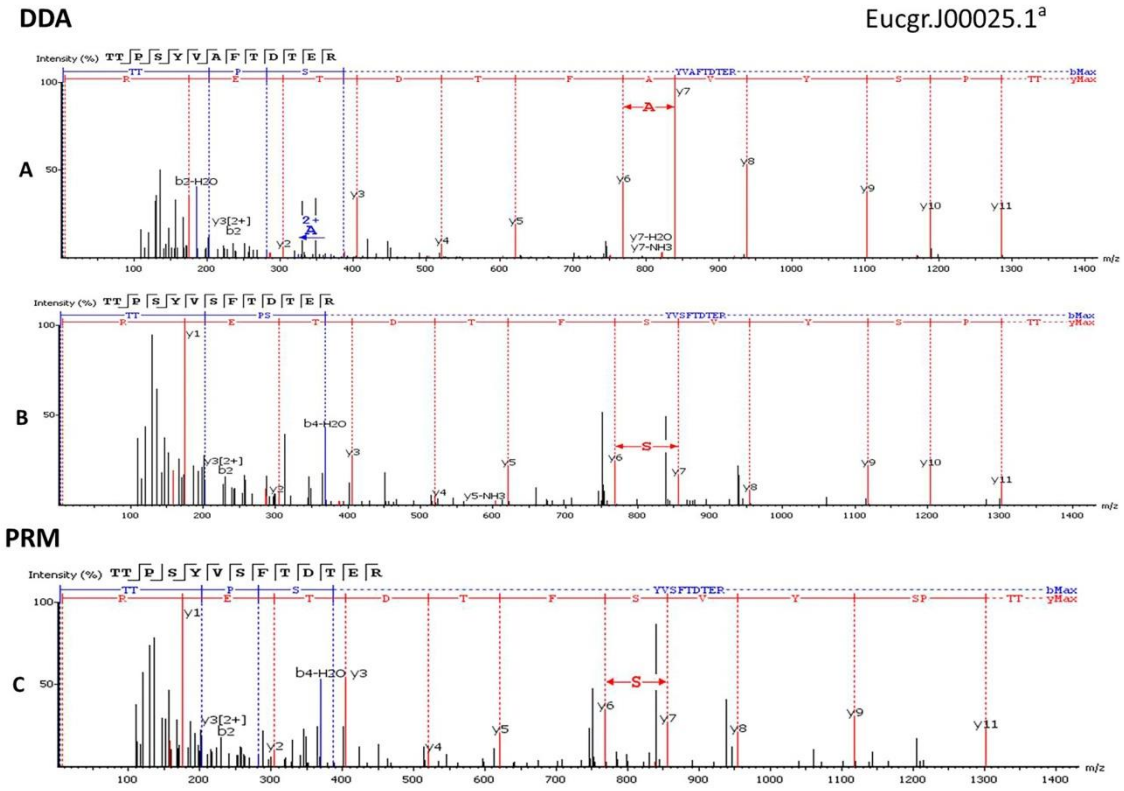


Fig. 4 – Tandem mass spectrometry (MS/MS) validation of a single amino acid substitution (SAS) feature. A: MS/MS data from native peptide acquired from the data-dependent acquisition (DDA) experiment; B: MS/MS data of a SAS peptide from DDA experiment; C: MS/MS data of a SAS peptide from the parallel reaction monitoring (PRM) experiment. [a] Phytozome Accession Code in which the single amino acid substituted peptide was identified according to the *E. grandis* v.2.0 reference protein database.

Parallel reaction monitoring experiments were also used to validate the existence of the 41 novel peptides (Table 2). The peptide retention times presented a minor shift between both experiments, and the mass-to-charge ratios of the precursor ion were closely the same, showing a very robust analytical performance. Additionally, mass-to-charge ratios of the three most intense product ions in both the DDA and PRM experiments were compared to confirm the existence of such peptides. This approach offered more accurate and reproducible evidence of the peptide identities, and provided an additional level of identification specificity.

Table 2 - Data-depend acquisition and parallel reaction monitoring results' comparison for the validation of 41 novel peptides identified by spectral correlations against modified RNA databases.

			Discovery-driven approach			Targeted approach		
Number	**Peptide Sequence**	**Charge [a]**	**Precursor m/z (RT) [b]**	**Product ion m/z [c]**	**PSM count [d]**	**Precursor m/z (RT) [e]**	**Product ion m/z [f]**	**PSM count [g]**
1	ADKNLLR	2	416.2518 (35.45)	175.1291 (y7); 288.2285 (y6); 401.3250 (y5)	6	416.2518 (34.62)	175.1294 (y7); 288.2243 (y6); 401.3250 (y5)	135
2	AGESLCYGVPAR	2	641.8113 (49.98)	129.1052 (b2); 174.0603 (b4(2+)); 442.3118 (b5)	1	641.8113 (49.79)	129.1063 (b2); 174.0606 (b4(2+)); 442.3128 (b5)	48
3	KLGLTLR	2	400.7731 (27.66)	175.1293 (y7); 288.1964 (y6); 541.4106 (y3)	3	400.7731 (27.36)	175.1280 (y7); 288.1974 (y6); 541.3600 (y3)	20
4	LSTMAAPTSGSGR	2	627.8097 (27.23)	177.1109 (y13); 376.2275 (y10); 463.2589 (y9)	7	627.8097 (27)	177.1109 (y13); 376.2276 (y10); 463.2588 (y9)	28
5	NHGIHFR	2	440.7321 (19.84)	175.1289 (y7); 322.2135 (y6); 459.2945 (y5)	1	440.7321 (19.79)	175.1290 (y7); 322.2136 (y6); 459.2945 (y5)	21
6	NLGSTANLR	2	473.2604 (23.21)	175.1287 (y9); 288.2238 (y8); 402.2803 (y7)	1	473.2604 (23.03)	175.1288 (y9); 288.2239 (y8); 402.2804 (y7)	41
7	NLGVGFAR	2	417.2362 (39.06)	175.1292 (y8); 450.2981 (y5); 606.4139 (y3)	9	417.2362 (38.66)	175.1293 (y8); 450.2989 (y5); 606.4137 (y3)	401
8	REVCMGWHHFHAIGQR	3	675.3237 (52.65)	175.1197 (y16); 965.5828 (y9); 868.3984 (y3(2+))	1	675.3237 (51.99)	175.1211 (y16); 965.6069 (y9); 868.4161 (y3(2+))	238
9	SAENIMMPR	2	526.2554 (38.58)	143.1236 (b2); 136.0815 (b3(2+)); 646.4163 (b6)	2	526.2554 (38.95)	143.1237 (b2); 136.0815 (b3(2+)); 646.4316 (b6)	207
10	VALSDPIR	2	435.7576 (32.82)	100.1150 (b1); 86.09855 (b2(2+)); 143.1245 (b3(2+))	6	435.7576 (32.11)	100.1160 (b1); 86.09896 (b2(2+)); 143.1258 (b3(2+))	102
11	VSLSIVPR	2	435.7755 (33.16)	175.1273 (y8); 272.1907 (y7); 571.4238 (y4)	4	435.7755 (34.12)	175.1291 (y8); 272.1909 (y7); 571.4239 (y4)	354
12	RSGGEMTTAAPR	2	617.3055 (27.88)	175.1270 (y12); 129.1057 (y11(2+)); 616.4299 (y7)	8	617.3055 (27.64)	175.1273 (y12); 129.1065 (y11(2+)); 616.4297 (y7)	60
13	EGLVSPLIK	2	478.2966 (24.8)	129.1072 (y9); 260.2153 (y8); 557.3976 (y5)	2	478.2966 (24.56)	129.1081 (y9); 260.2159 (y8); 557.4097 (y5)	45
14	SSTVKVTRK	2	503.3099 (22.47)	129.1068 (y9); 303.2274 (y8); 831.6248 (y3)	8	503.3099 (22.35)	129.1078 (y9); 303.2276 (y8); 831.6257 (y3)	84
15	VAVEVSPR	2	428.7495 (30.42)	86.09855 (b2(2+)); 136.0803 (b3(2+)); 201.1350 (b4(2+))	13	428.7495 (30)	86.09909 (b2(2+)); 136.0815 (b3(2+)); 201.1378 (b4(2+))	182
16	ARPFGAPR	2	437.7548 (31.21)	272.1909 (y7); 173.1391 (y6(2+)); 201.1330 (y5(2+))	6	437.7548 (31.33)	272.1910 (y7); 173.1392 (y6(2+)); 201.1353 (y5(2+))	86
17	RRNGVCLFYR	2	670.8513 (48.05)	175.1265 (y10); 598.3667 (y7); 457.7845 (y4(2+))	5	670.8513 (47.29)	175.1269 (y10); 598.3943 (y7); 457.7851 (y4(2+))	244
18	AASGNASLPR	2	473.2542 (30.57)	157.1393 (y10); 129.1071 (y9(2+)); 185.1475 (y8(2+))	15	473.2542 (31.83)	157.1413 (y10); 129.1081 (y9(2+)); 185.139 (y8(2+))	201
19	GSGQTNCCSSRNPRNGR	2	955.924 (51.83)	175.1246 (y17); 348.1258 (y15); 594.2767 (y8(2+))	3	955.924 (51.52)	175.1204 (y17); 348.1243 (y15); 594.2761 (y8(2+))	155
20	HIQIGLR	2	418.7603 (23.52)	175.1292 (y7); 586.4418 (y3); 699.5452 (y2)	6	418.7603 (22.33)	175.1293 (y7); 586.4432 (y3); 699.5482 (y2)	52
21	KPEDANR	2	416.215 (24.58)	175.1292 (y7); 145.0684 (y6(2+)); 475.2710 (y4)	1	416.215 (24.05)	175.1294 (y7); 145.0763 (y6(2+)); 475.2712 (y4)	5
22	LGTILSPR	2	428.7673 (31.61)	86.09864 (b2(2+)); 255.1652 (b3); 385.2910 (b4)	9	428.7673 (31.18)	86.09669 (b2(2+)); 255.1658 (b3); 385.2978 (b4)	673
23	MNLPVKR	2	437.2526 (49.42)	175.1290 (y7); 499.3595 (y4); 612.4504 (y3)	2	437.2526 (48.92)	175.1293 (y7); 499.3596 (y4); 612.4505 (y3)	94
24	RAIMRPR	2	450.2709 (32.13)	175.1289 (y7); 272.1906 (y6); 743.5468 (y2)	10	450.2709 (31.32)	175.1287 (y7); 272.1909 (y6); 743.5281 (y2)	211
25	KTAIEMSSMHEERK	2	846.903 (44.09)	129.1012 (y14); 304.1740 (y13); 847.5419 (y9)	2	846.903 (43.69)	129.1028 (y14); 304.178 (y13); 847.5137 (y9)	57
26	NLGLGFGR	2	417.2365 (40.09)	175.1292 (y8); 436.2794 (y5); 550.3508 (y4)	6	417.2362 (39.48)	175.1293 (y8); 436.2797 (y5); 550.3496 (y4)	361
27	APTTAPPAVR	2	491.7835 (24.65)	175.1284 (y10); 129.1071 (y9(2+)); 327.2245 (y8)	1	491.7835 (24.88)	175.1285 (y10); 129.1071 (y9(2+)); 327.2303 (y8)	43
28	KIGLTLR	2	400.7731 (27.66)	147.1197 (y8); 218.1640 (y7); 331.2585 (y6)	2	400.7731 (27.29)	147.1199 (y8); 218.1668 (y7); 331.2587 (y6)	27
29	KLVEDLAK	2	458.2809 (20.26)	130.0907 (y7(2+)); 361.2708 (y6); 508.3597 (y5)	3	458.2809 (20.15)	130.0908 (y7(2+)); 361.2769 (y6); 508.3702 (y5)	7
30	LFPLFSVR	2	490.796 (66.93)	175.1279 (y8); 273.1384 (y7); 707.4218 (y4)	2	490.796 (65.74)	175.1281 (y8); 273.1385 (y7); 707.4045 (y4)	60
31	LLCCQFPR	2	547.27 (20.28)	175.1246 (y15); 272.1837 (y14); 887.4899 (y9)	4	547.27 (20.4)	175.1254 (y15); 272.1902 (y14); 887.4888 (y9)	29
32	QGKPTCPDHESTCPR	2	886.8912 (27.07)	175.1280 (y10); 290.2057 (y9); 501.3660 (y7)	1	886.8912 (26.7)	175.1281 (y10); 290.2099 (y9); 501.3654 (y7)	63
33	SVAIGERGLR	2	530.813 (56.98)	175.1281 (y9); 432.2953 (y7); 618.4004 (y6)	9	530.813 (56.4)	175.128 (y9); 432.2955 (y7); 618.3859 (y6)	432
34	TVVGLWPCR	2	544.2927 (37.37)	86.9864 (b2(2+)); 255.1652 (b3); 472.3413 (b5)	6	544.2927 (36.5)	86.9869 (b2(2+)); 255.1665 (b3); 472.3364 (b5)	244
35	IGTVTIPR	2	428.7673 (31.61)	175.1289 (y8); 385.2910 (y6); 486.3239 (y5)	16	428.7673 (30.78)	175.1291 (y8); 385.2913 (y6); 486.3283 (y5)	539
36	PASSSRRPSR	2	551.8009 (26.48)	175.1278 (y10); 758.4974 (y5); 845.5580 (y4)	4	551.8009 (26.85)	175.1279 (y10); 758.5169 (y5); 845.559 (y4)	47
37	TVERLSMVEDVR	2	726.8776 (54.64)	185.1663 (b2); 470.3186 (b4); 932.6211 (b8)	2	726.8776 (54.32)	185.1681 (b2); 470.3177 (b4); 932.6211 (b8)	211
38	VTAAAADASPSR	2	558.7863 (23.1)	129.1046 (b3(2+)); 414.2797 (b5); 485.3318 (b6)	10	558.7863 (22.81)	129.1071 (b3(2+)); 414.2793 (b5); 485.3319 (b6)	52
39	APLALSPR	2	412.7536 (31.05)	175.1312 (y8); 128.0879 (y7(2+)); 472.3461 (y5)	9	412.7536 (31.3)	175.1293 (y8); 128.0880 (y7(2+)); 472.3456 (y5)	277
40	EGDVLTLLESER	2	681.3566 (66.84)	175.1204 (y12); 391.2113 (y10); 633.3661 (y8)	7	681.3566 (66.2)	175.1267 (y12); 391.2101 (y10); 633.3522 (y8)	356
41	IPTGATLKGPMMTSPVLVEMAR	3	784.0708 (58.26)	175.1159 (y22); 197.1268 (y20(2+)); 522.2659 (y19)	3	784.0708 (57.97)	175.1156 (y22); 197.1268 (y20(2+)); 522.3196 (y19)	112

[a]charge state of the precursor ion;[b, e] mass-to-charge ratios and best spectrum retention times of the precursor ion from data-depend acquisition (DDA) experiment and parallel reaction monitoring (PRM) experiment, respectively; [c, f] mass-to-charge ratios of three most intense product ions from DDA and PRM experiments, respectively; [d, g] Number of peptide spectrum matches (PSM) in the ".sqt" file from DDA and PRM experiments, respectively.

4. DISCUSSION

Identification of new peptide forms (not present in a reference protein sequence database) through proteogenomics is a powerful approach, as it provides new protein-level evidence of gene expression. In the present study, novel peptide forms, as well as SASs, were confidently identified in *E. grandis* through a dedicated proteogenomics workflow. By using the pipeline proposed herein, it was possible to conducte a more comprehensive genomic characterization of novel peptide forms derived from both known and novel coding *loci*. In the present manuscript, we specifically aimed at showing the potential of the proteogenomics approach in stem cells. Proteome analysis of stem tissues consistently result in the identification of low-abundant proteins/peptides since it has fewer interfering compounds and lower proteome dynamic range than leaf tissues.

The genomic mapping strategy is relevant as it provides extra information on the specific type of identified peptides in proteogenomic studies. Most novel peptides were mapped to exon regions as out of frame peptides (EXON-ARF). It is worth mentioning that this class of novel peptides was only considered if the identified reading frame was different from the longest reading frame predicted for the same *locus*. The second most abundant genomic classification consisted of peptides mapped to novel coding *loci*, which had previous RefSeq annotations as non-protein coding regions. Furthermore, peptides mapped to untranslated regions (3′ or 5′ UTR peptides) and those spanning the boundary between the coding sequence region and the neighboring 3’ UTR, 5’ UTR or intron regions were also identified. Nevertheless, alternative splicing events could not be identified because of limitations to the modified databases used in this study since the order of exon-exon junctions was maintained in the new theoretical protein sequences created in the chimeric databases.

The same novel peptide sequence may be originated from multiple genomic locations (e.g., gene paralogues, or a protein-coding gene and a pseudogene) (Nesvizhskii 2014). In this study, four novel peptides were mapped to two different genomic locations, which demonstrate that these events might have resulted from gene duplication events. Moreover, a novel peptide mapped to a single genomic location may also have multiple interpretations. For instance, the same peptide can

show two genomic classifications regarding two different transcripts of the same gene (Nesvizhskii 2014). Subsequently, approximately 70% of the novel peptides identified could be classified as orthologous, since they presented significant hits (mostly to the bacterium, fungus, plant and animal kingdoms, respectively (data not shown)) through *blastp* search against the NCBI non-redundant protein database.

Among some of the novel peptides, translational stop codon readthrough events were also detected. This mechanism was first identified in viruses, and it is used to increase the expression of viral proteins while preserving a compact genome size (Dreher and Miller 2006; Firth and Brierly 2012). This strategy has also been recognized in some eukaryotic organisms, such as *Drosophila* and other metazoa (Jungrei *et al.* 2011), as well as in plants (Liu 2004). Some readthrough events may be attributed to nonstop mutations, also named stop-loss mutations, which can change a stop codon into a sense codon, leading to translation into the 3′ untranslated region (3′ UTR) (Dhamija *et al.* 2020).

A *tblastn* search against *E. grandis* expressed sequence tag (EST) sequences provided insights into the existence of seven cDNA molecules. Two out of seven peptides, according to previous RefSeq database annotations, are derived from predicted ncRNA transcripts, which highlight that only RNA molecules were expected for these *loci*, and not necessarily, the existence of a protein originated from those RNA sequences. Therefore, this approach provides additional evidence to support the existence of such RNA sequences effectively expressed by the plant, and can improve gene annotation models. Proteogenomics has similarly contributed to improving genome annotation in other plants such as grapevine (Chapman and Bellgard 2017).

Single amino acid substitution, which usually cannot be identified in reference protein database searches, is an important source of variability that could enhance plant adaptability. Analysis of shotgun proteomics data generally relies on search in databases that usually do not include information on protein variants, and thus prevents SASs from being identified (Li *et al.* 2011). Through a homology search based on *de novo* sequencing, the identity of six peptides containing at least one SAS could be directly inferred. The amino acid composition of each protein can be largely influenced by indirect selection pressures (evolutionary forces acting upon random genetic mutations) (Abraham *et al.* 2015). A slight change in the amino acid

composition may alter the activity, shape, and/or function of a protein, which can a) have no impact on its functionality; b) effectively improve its functionality; c) change its function; d) be deleterious, changing the protein structure or function (Abraham *et al.* 2015). However, defining the functional impact of each variant is beyond the scope of this research.

The PRM experiment was used as a validation method. PRM is an ion monitoring technique based on high-resolution and high-precision mass spectrometry that can target specific ions of interest. In this study, the validation success was 100% over the novel identifications, showing that PRM was a highly sensitive method. Thus, the results accomplished by the DDA experiment agreed with those of the PRM method for both novel peptides and SASs. Confirmation of novel peptides by independent targeted proteomics has also been carried out in previous studies using PRM assays (Omasits *et al.* 2017).

In brief, the present study demonstrates that novel peptide forms and SASs can be confidently identified in the *Eucalyptus grandis* proteome through a dedicated proteogenomics workflow, which is composed of both spectral correlations against *in silico* modified RefSeq databases and a *de novo* sequencing technique. This approach can assist with expanding the growing field of proteogenomics in plant science, and has the potential to provide novel insights into biological regulation mechanisms as new peptide forms are discovered, as well as to refine gene annotation models.

5. DATA ARCHIVING STATEMENT

The mass spectrometry proteomics (RAW files and list of protein IDs) data have been deposited in the ProteomeXchange Consortium via the PRIDE (Perez-Riverol *et al.* 2019) partner repository with dataset identifier PXD022110.

Funding

This study was supported by grant no. 400459/2016-7 from the National Council for Scientific and Technological Development (CNPq) and grant no. 2018/15035-8 from the São Paulo Research Foundation (FAPESP).

Acknowledgments

T.S.B. thanks CNPq for the research fellowship. G.L.J. thanks CNPq for the PhD scholarship.

Conflict of Interest

The authors declare having no conflicts of interest that could be perceived as prejudicial to the impartiality of the reported research.

Author Contributions

Conceptualization: [Tiago Santana Balbuena]; Methodology: [Gabriel Lemes Jorge, Tiago Santana Balbuena]; Formal analysis and investigation: [Gabriel Lemes Jorge, Tiago Santana Balbuena], Writing - original draft preparation: [Gabriel Lemes Jorge]; Writing - review and editing: [Tiago Santana Balbuena, Gabriel Lemes Jorge]; Funding acquisition: [Tiago Santana Balbuena]; Resources: [Tiago Santana Balbuena]; Supervision: [Tiago Santana Balbuena].

Consent for publication

All authors have approved the final version of the manuscript for publication.

Availability of data and material (Not applicable)

Code availability (Not applicable)

Ethics approval (Not applicable)

Consent to participate (Not applicable)

6. REFERENCES

Abraham PE, Wang X, Ranjan P, Nookaew I, Zhang B, Tuskan GA, Hettich RL (2015) Integrating mRNA and Protein Sequencing Enables the Detection and Quantitative Profiling of Natural Protein Sequence Variants of *Populus trichocarpa*. **J Proteome Res** 14:5318-5326.

Ang MY, Low TY, Lee PY, Wan Mohamad Nazarie WF, Guryev V, Jamal R (2019) Proteogenomics: From next-generation sequencing (NGS) and mass spectrometry-based proteomics to precision medicine. **ClinChimActa** 498:38-46.

Baerenfaller K, Grossmann J, Grobei MA, Hull R, Hirsch-Hoffmann M, Yalovsky S, Zimmermann P, Grossniklaus U, Gruissem W, Baginsky S (2008) Genome-scale proteomics reveals *Arabidopsis thaliana* gene models and proteome dynamics. **Science** 320:938-941.

Blank-Landeshammer B, Kollipara L, Biß K, Pfenninger M, Malchow S, Shuvaev K, Zahedi RP, Sickmann A (2017) Combining de novo peptide sequencing algorithms, a synergistic approach to boost both identifications and confidence in bottom-up proteomics. **J Proteome Res** 16(9):3209-3218.

Blank-Landeshammer B, TeichertI, Märker R, Nowrousian M, Kück U, Sickmann A (2019) Combination of proteogenomics with peptide de novo sequencing identifies new genes and hidden posttranscriptional modifications. **mBio** 10:e02367-e02319.

Carvalho PC, Fischer JS, Chen EI, Yates JR, Barbosa VC (2008) PatternLab for proteomics: a tool for differential shotgun proteomics. **BMC Bioinformatics** 21:316.

Carvalho PC, Fischer JS, Xu T, Yates JR, Barbosa VC (2012) PatternLab: from mass spectra to label-free differential shotgun proteomics. **CurrProtocBioinformatics** 13: 13-19.

Carvalho PC, Lima DB, Leprevost FV, Santos MD, Fischer JS, Aquino PF, Moresco JJ, Yates JR, Barbosa VC (2016) Integrated analysis of shotgun proteomic data with PatternLab for proteomics 4.0. **Nat Protoc** 11:102-117.

Castellana NE, Payne SH, Shen Z, Stanke M, Bafna V, Briggs SP (2008) Discovery and revision of *Arabidopsis* genes by proteogenomics. **PNAS** 105(52):21034-21038.

Chapman B,Bellgard M (2017) Plant Proteogenomics: Improvements to the grapevine genome annotation. **Proteomics** 17:1700197.

Dhamija S, Yang CM, Seiler J, Myacheva K, Caudron-Herger M, Wieland A, Abdelkarim M, Sharma Y, Riester M, Groß M, et al (2020) A pan-cancer analysis reveals nonstop extension mutations causing SMAD4 tumour suppressor degradation. **Nat Cell Biol** 22:999-1010.

Dreher TW, Miller WA (2006) Translational control in positive strand RNA plant viruses. **Virology** 344:185–197.

Eng J, McCormack AL, Yates JR (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. **J Am Soc Mass Spectrom** 5:976-989.

Firth AE,Brierley I (2012) Non-canonical translation in RNA viruses. **J Gen Virol** 93:1385–1409.

Gao X, Li H, Li H, Dong S, Chu J, Guo H, Zhao Q (2018) Sensitive determination of nine anticoagulant rodenticides in blood by high resolution mass spectrometry with supported liquid extraction pretreatment. **Forensic SciInt** 292:39-44.

Han Y, Ma B, Zhang K (2005) SPIDER: Software for Protein Identification from Sequence Tags Containing De Novo Sequencing Error. **J BioinformComput Biol** 3(3): 697-716.

Heunis T, Dippenaar A, Warren RM, Van Helden PD, Van der Merwe RG, Gey Van Pittius NC, Pain A, Sampson SL, Tabb DL (2017) Proteogenomic investigation of strain variation in clinical *Mycobacterium tuberculosis* isolates. **J Proteome Res** 16(10):3841–51.

Jaffe JD, Berg HC, Church GM (2004) Proteogenomic mapping as a complementary method to perform genome annotation. **Proteomics** 4:59-77.

Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, Mende DR, Letunic I, Rattei T, Jensen LJ, Mering CV, Bork P (2019) EggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. **Nucleic Acids Res** 47: D309–D314.

Jungreis I, Lin MF, Spokony R, Chan CS, Negre N, Victorsen A, White KP, Kellis M (2011) Evidence of abundant stop codon readthrough in *Drosophila* and other metazoa. **Genome Res** 21:2096–2113.

Li J, Su Z, Ma Ze-Q, Slebos RJC, Halvey P, Tabb DL, Liebler DC, Pao W, Zhang B (2011) A bioinformatics workflow for variant peptide detection in shotgun proteomics. **Mol Cell Proteomics** 10:M110.006536.

Liu Q, Xue Q (2004) Computational identification and sequence analysis of stop codon readthrough genes in *Oryza sativa*. **Biosystems** 77:33–39.

Myburg A, Grattapaglia D, Tuskan G, Hellsten U, Hayes RD, Grimwood J, Jenkins J, Lindquist E, Tice H, Bauer D et al (2014) The genome of *Eucalyptus grandis*. **Nature** 510:356-362.

Nesvizhskii, A.I (2014) Proteogenomics: concepts, applications and computational strategies. **Nat. Methods** 11:1114–1125.

Omasits U, Varadarajan AR, Schmid M, Goetze S, Melidis D, Bourqui M, Nikolayeva O, Québatte M, Patrignani A, Dehio C, et al (2017) An integrative strategy to identify the entire protein coding potential of prokaryotic genomes by proteogenomics. **Genome Res** 27:2083–2095.

Peterson AC, Russell JD, Bailey DJ, Westphall MS, Coon JJ (2012) Parallel reaction monitoring for high resolution and high mass accuracy quantitative, targeted proteomics. **Mol Cell Proteomics** 11:1475–1488.

Perez-Riverol Y, Csordas A, Bai J, Bernal-Llinares M, Hewapathirana S, Kundu DJ, Inuganti A, Griss J, Mayer G, Eisenacher M, et al (2019) The PRIDE database and related tools and resources in 2019: improving support for quantification data. **Nucleic Acids Res** 47(D1): D442-D450.

Potgieter MG, Nakedi KC, Ambler JM, Nel AJM, Garnett S, Soares NC, Mulder N, Blackburn JM (2016) Proteogenomic analysis of *Mycobacterium smegmatis* using high resolution mass spectrometry. **Front Microbiol** 7:427.

Rauniyar N (2015) Parallel reaction monitoring: a targeted experiment performed using high resolution and high mass accuracy mass spectrometry. **Int J MolSci** 16: 28566-28581.

Sandow JJ, Rainczuk A, Infusini G, Makanji M, Bilandzic M, Wilson AL, Fairweather N, Stanton PG, Garama D, Gough D *et al.* (2018) Discovery and validation of novel protein biomarkers in ovarian cancer patient urine. **Proteomics ClinAppl** 12(3):1700135.

Shevchenko A, Tomas H, Havlis J, Olsen JV, Mann M (2007) In-gel digestion for mass spectrometric characterization of proteins and proteomes. **Nat Protoc** 1:2856-2860.

Wernersson R (2006) Virtual Ribosome - a comprehensive translation tool with support for sequence feature integration. **Nucl Acids Res** 34:385-388.

Zhang J, Xin L, Shan B, Chen W, Xie M, Yuen D, Zhang W, Zhang Z, Lajoie AG, Ma B (2011) PEAKS DB: *de novo* sequencing assisted database search for sensitive and accurate peptide identification. **Mol. Cell. Proteomics** 11(4):M111.010587.

Supplementary material

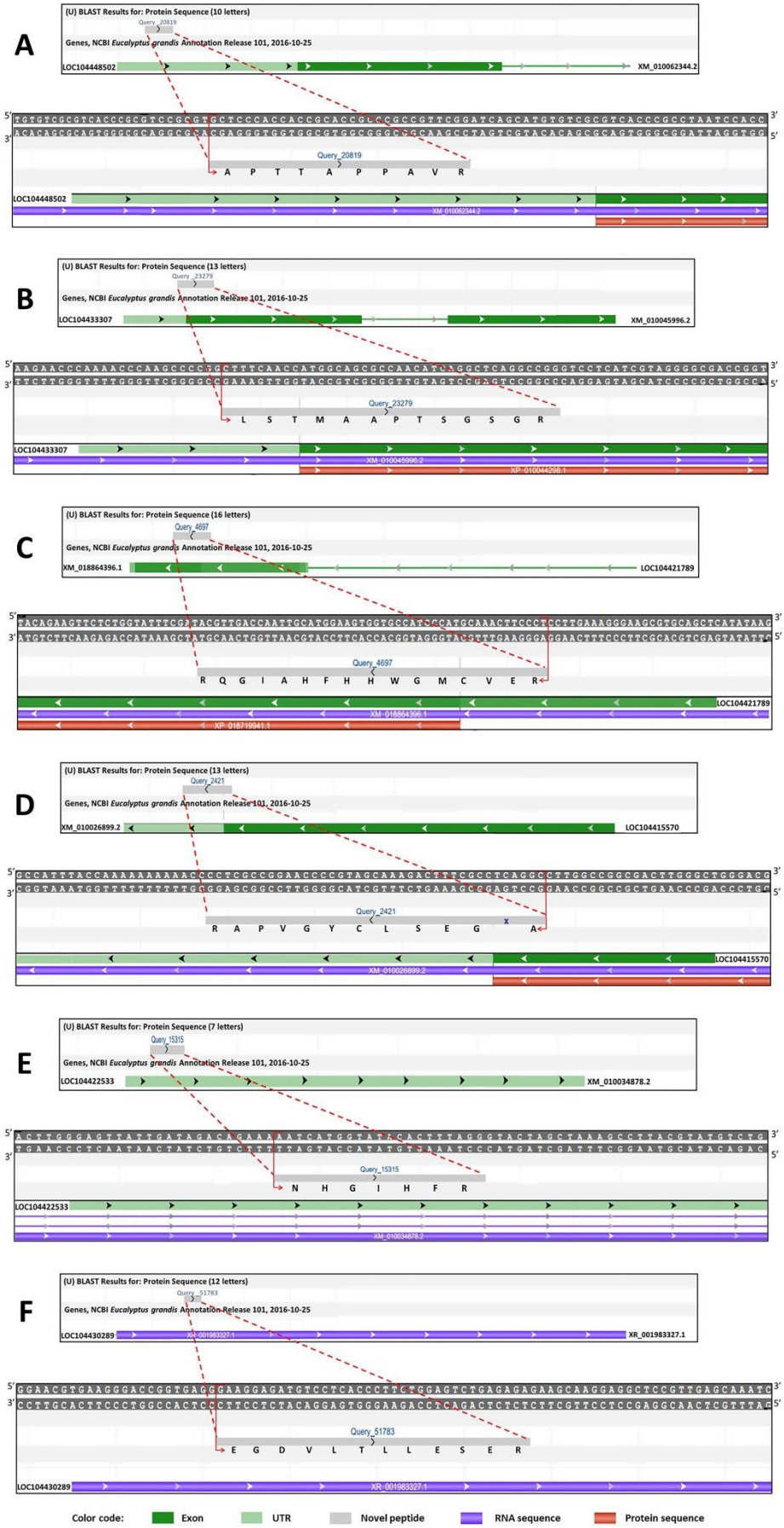

A
(U) BLAST Results for: Protein Sequence (10 letters)
Genes, NCBI Eucalyptus grandis Annotation Release 101, 2016-10-25
LOC104448502
XM_010062344.2
Query_20819
A P T T A P P A V R
B
(U) BLAST Results for: Protein Sequence (13 letters)
Genes, NCBI Eucalyptus grandis Annotation Release 101, 2016-10-25
LOC104433307
XM_010045996.2
Query_23279
L S T M A A P T S G S G R
XP_010044298.1
C
(U) BLAST Results for: Protein Sequence (16 letters)
Genes, NCBI Eucalyptus grandis Annotation Release 101, 2016-10-25
XM_018864396.1
LOC104421789
Query_4697
R Q G I A H F H H W G M C V E R
XP_018719941.1
D
(U) BLAST Results for: Protein Sequence (13 letters)
Genes, NCBI Eucalyptus grandis Annotation Release 101, 2016-10-25
XM_010026899.2
LOC104415570
Query_2421
R A P V G Y C L S E G A
E
(U) BLAST Results for: Protein Sequence (7 letters)
Genes, NCBI Eucalyptus grandis Annotation Release 101, 2016-10-25
LOC104422533
XM_010034878.2
Query_15315
N H G I H F R
F
(U) BLAST Results for: Protein Sequence (12 letters)
Genes, NCBI Eucalyptus grandis Annotation Release 101, 2016-10-25
LOC104430289
XR_001983327.1
Query_51783
E G D V L T L L E S E R
Color code:
Exon
UTR
Novel peptide
RNA sequence
Protein sequence

Fig. S1 – Mapping illustrations of the genomic classifications of the novel peptides. (A) Peptide mapped to a 5′ untranslated region (5’ UTR); (B) peptide mapped to regions between the coding sequence and the neighboring 5’ UTR region (Exon boundary) (5’ UTR–EXON); (C) peptide mapped to regions between the coding sequence and the neighboring intron region (Exon boundary) (INTRON–EXON); (D) peptide mapped to regions between the coding sequence and the neighboring 3’ UTR region (Exon boundary) (EXON–3’ UTR); (E) peptide mapped to a 3′ untranslated region (3’ UTR); (F) peptide mapped to a novel coding locus.

Table S1 – Genomic assignment data used for the novel peptide mapping against the *E. grandis* genome (Egrandis1_0 reference Annotation Release 101).

Peptide number	Pepetide sequence	RNA accession number[a]	Longest protein RF[b]	Novel peptide RF[c]	Genomic classification	Expect threshold	Max score	Query coverage	Identity
1	ADKNLLR	XM_010055119.2	2	1	EXON - ARF	10000	17.3	100	100
2	A*GESLCYGVPAR	XM_010026899.2	3	1	EXON - 3'UTR	100	26.9	100	100
3	KLG*LT*LR	XM_010037706.2	3	1	EXON - ARF	10000	16.9	100	100
4	LSTMAAPTSGSGR	XM_010045996.2	1	1	5' UTR - EXON	100	27.7	100	100
5	NHGIHFR	XM_010034878.2	1	1	3'UTR	1000	19.2	100	100
6	NLGSTANLR	XM_018860691.1	2	1	5' UTR	1000	20.4	100	100
7	NLGVGFAR	XM_010053962.2	1	1	5' UTR	100	21.2	100	100
8	REVCMGWHHFHAIGQR	XM_018864396.1	1	1	INTRON -EXON	100	38.5	100	100
9	SAENIMMPR	XM_010037390.2	3	1	EXON - ARF	100	22.3	100	100
10	VALSDPIR	XM_018865801.1	1	1	5' UTR	1000	18.9	100	100
11	VSLSIVPR	XM_010027007.2	3	1	EXON - ARF	1000	18.5	100	100
12	RSGGEMTTAAPR	XM_010068985.2 / XR_727223.2	1 / 1	1 / 1	5' UTR - EXON / NCL	100	26.9	100	100
13	EGLVSPLIK	XM_018869853.1	3	1	EXON - ARF	1000	20.8	100	100
14	SSTVKVTRK	XM_010062006.2	2	1	INTRON -EXON	10000	16.2	100	86
15	VAVEVSPR	XM_018862919.1	3	1	5' UTR	1000	18.5	100	100
16	ARPFGAPR	XM_018859771.1 / XR_001979892.1	2 / 2	1 / 1	EXON - ARF / NCL	1000	19.2	100	100
17	RRNGVCLFYR	XM_018861651.1 / XR_001980976.1	2 / 3	1 / 2	EXON - ARF / NCL	1000	25	100	100
18	AASGNASLPR	XM_018861733.1	2	2	5' UTR	100	21.6	100	100
19	GSGQTNCCSSRNPRNGR	XM_010046998.2	3	2	EXON - ARF	100	37	100	100
20	HIQIGLR	XM_010039327.2	1	2	EXON - ARF	10000	18.1	100	100
21	KPEDANR	XM_010045581.2	1	2	EXON - ARF	1000	18.1	100	100
22	LGTI**LSPR	XM_010065691.2	1	2	5' UTR	1000	20	100	100
23	MNLPVKR	XM_018873041.1	1	2	EXON - ARF	10000	17.7	100	100
24	RAIMRPR	XM_010029799.2	1	2	EXON - ARF	10000	16.9	100	100
25	KTAIEMSSMHEERK	XM_010054572.2	3	2	EXON - ARF	100	30.8	100	100
26	NLGLGFGR	XM_018876096.1	3	2	5' UTR	1000	19.6	100	100
27	APTTAPPAVR	XM_010062344.2	2	3	5' UTR	100	21.6	100	100
28	KIGLTLR	XM_010069929.1	1	3	EXON - ARF	1000	18.1	100	100
29	KLV*ED*LAK	XM_010046007.2 / XM_018872208.1	2 / 3	3 / 3	EXON - ARF / 3‘ UTR	1000	18.9	100	100
30	LFPLFSVR	XM_018870367.1	1	3	EXON - ARF	1000	19.6	100	100
31	LL*CC**QFPR	XM_018870037.1	1	3	EXON - ARF	100	21.6	100	100
32	QGKPTCPDHESTCPR	XM_010070621.1	1	3	EXON - ARF	100	35	100	100
33	SVAIGERGLR	XM_010044558.2	1	3	EXON - ARF	100	22.7	100	100
34	TVVGLW*PCR	XM_018875477.1	1	3	EXON -ARF	100	23.5	100	100
35	IGTVTIPR	XM_018871868.1	1	3	DUBIOUS	-	-	-	-
36	PASSSRRPSR	XR_720815.2	2	1	NCL	1000	20	100	100
37	TVERLSMVEDVR	XR_001984062.1	2	2	NCL	100	26.2	100	100
38	VTAAAADASPSR	XR_001979828.1	1	2	NCL	100	24.3	100	100
39	APLALSPR	XR_721165.2	3	1	NCL	10000	17.7	100	100
40	EGDVLTLLESER	XR_001983327.1	2	2	NCL	100	26.6	100	100
41	IPTGATLKGPMMTSPVLVEMAR	XR_001981593.1	1	2	NCL	100	45.8	100	100

*Indicates the presence of a stop codon within the novel peptide sequence; [a] only the most representative NCBI accession number is listed for each genomic classification; [b,c] the longest protein reading frame and novel peptide reading frame identification, at each NCBI protein locus, respectively; NCL: novel coding locus.

Table S2 – Characterization of novel peptides for their presence in previous cDNA libraries accomplished by a tblastn search against *E. grandis* EST db.

Pep.	Sequence	Max Score	Coverage	Identity	Accession[a]	Description
1	ADKNLLR	15.8	100%	100%	HS051946.1	GR-SE-001-GO-071-D12 Seedlings (EUGR-SE) EucalyptusgrandiscDNA clone GR-SE-001-GO-071-D12 5', mRNAsequence
2	A*GESLCYGVPAR	20.4	100%	67%	HS071885.1	GR-XY-006-RS-044-A03 Xylem (EUGR-XY) EucalyptusgrandiscDNA clone GR-XY-006-RS-044-A03 5', mRNAsequence
3	KLG*LT*LR	16.5	100%	86%	HS057449.1	GR-SE-001-RS-008-E09 Seedlings (EUGR-SE) EucalyptusgrandiscDNA clone GR-SE-001-RS-008-E09 5', mRNAsequence
4	LSTMAAPTSGSGR	28.1	100%	100%	HS041823.1	GR-PU-000-MG-018-A06.R EucalyptusgrandiscDNA clone GR-PU-000-MG-018-A06.R 5', mRNAsequence
5	NHGIHFR	19.2	92%	50%	HS048098.1	GR-SE-001-AF-017-C10 Seedlings (EUGR-SE) EucalyptusgrandiscDNA clone GR-SE-0
6	NLGSTANLR	18.1	100%	78%	HS044832.1	GR-PU-003-MG-014-B11 Young leaves infected by Pucciniapsidii (EUGR-PU) Eucalyptus
7	NLGVGFAR	18.5	100%	88%	HS067380.1	GR-TS-002-CN-071-B04 Treated seedlings (EUGR-TS) Eucalyptus grandiscDNA clone G
8	REVCMGWHHFHAIGQR	25.8	100%	50%	HS057129.1	GR-SE-001-RS-003-B02 Seedlings (EUGR-SE) EucalyptusgrandiscDNA clone GR-SE-0
9	SAENIMMPR	16.5	88%	75%	HS066183.1	GR-TS-002-CN-050-D04 Treated seedlings (EUGR-TS) Eucalyptus grandiscDNA clone G
10	VALSDPIR	17.7	87%	86%	HS054839.1	GR-SE-001-GO-111-G01 Seedlings (EUGR-SE) EucalyptusgrandiscDNA clone GR-SE
11	VSLSIVPR	17.3	100%	63%	HO765256.1	GR-ML-001-BA-011-A12 Mature leaves (EUGR-ML) EucalyptusgrandiscDNA clone GR
12	RSGGEMTTAAPR	18.5	100%	50%	HS045175.1	GR-PU-003-MG-019-A03 EucalyptusgrandiscDNA clone GR-PU-003-MG-019-A03 5', mRNAsequence
13	EGLVSPLIK	17.7	88%	75%	HS070419.1	GR-XY-006-RS-017-A03 Xylem (EUGR-XY) EucalyptusgrandiscDNA clone GR-XY-006-RS-017-A03 5', mRNAsequence
14	SSTVKVTRK	15.8	100%	56%	HS043297.1	GR-PU-003-GO-071-A01.F EucalyptusgrandiscDNA clone GR-PU-003-GO-071-A01.F 5', mRNAsequence
15	VAVEVSPR	16.5	87%	86%	HS055838.1	GR-SE-001-GO-134-H03 Seedlings (EUGR-SE) EucalyptusgrandiscDNA clone GR-SE-001-GO-134-H03 5', mRNAsequence
16	ARPFGAPR	19.2	100%	88%	HS065919.1	GR-TS-002-CN-045-F08 Treated seedlings (EUGR-TS) Eucalyptus grandiscDNA clone
17	RRNGVCLFYR	25.4	90%	100%	HS069205.1	GR-TS-003-GO-014-D10 Treated seedlings (EUGR-TS) Eucalyptus grandiscDNA clone
18	AASGNASLPR	20.0	100%	90%	HS072251.1	GR-XY-006-RS-054-C11 Xylem (EUGR-XY) Eucalyptus grandiscDNA clone GR-XY-006
19	GSGQTNCCSSRNPRNGR	37.0	100%	100%	HS067133.1	GR-TS-002-CN-067-A10 Treated seedlings (EUGR-TS) Eucalyptus grandiscDNA clone
20	HIQIGLR	16.2	100%	86%	HS055552.1	GR-SE-001-GO-127-H08 Seedlings (EUGR-SE) EucalyptusgrandiscDNA clone GR-SE
21	KPEDANR	16.5	85%	83%	HO769082.1	GR-ML-001-BA-072-B09 Mature leaves (EUGR-ML) EucalyptusgrandiscDNA clone GR-ML-001-BA-072-B09 5', mRNAsequence
22	LGTI**LSPR	17.3	100%	75%	HS045861.1	GR-PU-003-MG-029-G12EucalyptusgrandiscDNA clone GR-PU-003-MG-029-G12 5', mRNAsequence
23	MNLPVKR	17.3	100%	86%	HS070418.1	GR-XY-006-RS-017-A04 Xylem (EUGR-XY) EucalyptusgrandiscDNA clone GR-XY-006-RS-017-A04 5', mRNAsequence
24	RAIMRPR	15.8	85%	67%	HS062551.1	GR-TS-001-GO-055-B04 Treated seedlings (EUGR-TS) Eucalyptus grandiscDNA clone G
25	KTAIEMSSMHEERK	19.2	100%	50%	HS073332.1	GR-XY-006-RS-075-A07.R Xylem (EUGR-XY) EucalyptusgrandiscDNA clone GR-XY-006-RS-075-A07.R 5', mRNAsequence
26	NLGLGFGR	19.2	100%	100%	HS069330.1	GR-TS-003-GO-018-D11 Treated seedlings (EUGR-TS) Eucalyptus grandiscDNA clone
27	APTTAPPAVR	17.7	90%	78%	HS057812.1	GR-SE-002-CN-002-A05 Seedlings (EUGR-SE) EucalyptusgrandiscDNA clone GR-SE
28	KIGLTLR	15.4	100%	71%	HS057449.	GR-SE-001-RS-008-E09 Seedlings (EUGR-SE) EucalyptusgrandiscDNA clone GR-SE
29	KLV*ED*LAK	16.2	100%	75%	HS066341.1	GR-TS-002-CN-053-E10 Treated seedlings (EUGR-TS) Eucalyptus grandiscDNA clone GR-TS-002-CN-053-E10 5', mRNA sequence
30	LFPLFSVR	17.7	87%	71%	HS059208.1	GR-TS-001-AF-012-C08 Treated seedlings (EUGR-TS) Eucalyptus grandiscDNA clone GR-TS-001-AF-012-C08 5', mRNA sequence
31	LL*CC**QFPR	19.2	87%	86%	HS071109.1	GR-XY-006-RS-029-D10 Xylem (EUGR-XY) EucalyptusgrandiscDNA clone GR-XY-006-RS-029-D10 5', mRNAsequence
32	QGKPTCPDHESTCPR	20.4	80%	58%	HS067872.1	GR-TS-002-RS-003-B01 Treated seedlings (EUGR-TS) Eucalyptus grandiscDNA clone GR-TS-002-RS-003-B01 5', mRNA sequence
33	SVAIGERGLR	22.7	100%	100%	HS066203.1	GR-TS-002-CN-051-D11 Treated seedlings (EUGR-TS) Eucalyptus grandiscDNA clone G
34	TVVGLW*PCR	18.9	100%	67%	HS061782.1	GR-TS-001-GO-044-B06 Treated seedlings (EUGR-TS) Eucalyptus grandiscDNA clone GR-TS-001-GO-044-B06 5', mRNA sequence
35	IGTVTIPR	16.9	87%	86%	HS061104.1	GR-TS-001-AF-037-H08.R Treatedseedlings (EUGR-TS) EucalyptusgrandiscDNA clone GR-TS-001-AF-037-H08.R 5', mRNAsequence
36	PASSSRRPSR	20.0	100%	100%	HS060478.1	GR-TS-001-AF-029-G03.R Treated seedlings (EUGR-TS) Eucalyptus grandiscDNA clone GR-TS-001-AF-029-G03.R 5', mRNA sequence
37	TVERLSMVEDVR	17.3	75%	78%	HS052698.1	GR-SE-001-GO-081-H07 Seedlings (EUGR-SE) EucalyptusgrandiscDNA clone GR-SE-001-GO-081-H07 5', mRNAsequence
38	VTAAAADASPSR	17.3	100%	58%	HS040940.1	GR-PU-000-MG-004-D07.R Young leaves infected by Pucciniapsidii (EUGR-PU) Eucalyp
39	APLALSPR	18.5	87%	86%	HS056745.1	GR-SE-001-GO-160-C12 Seedlings (EUGR-SE) EucalyptusgrandiscDNA clone GR-SE-001-GO-160-C12 5', mRNAsequence
40	EGDVLTLLESER	26.6	100%	100%	HO765756.1	GR-ML-001-BA-018-E06 Mature leaves (EUGR-ML) EucalyptusgrandiscDNA clone GR-ML-001-BA-018-E06 5', mRNAsequence
41	IPTGATLKGPMMTSPVLVEMAR	37.0	95%	81%	HO768082.1	GR-ML-001-BA-057-A11 Mature leaves (EUGR-ML) EucalyptusgrandiscDNA clone GR-ML-001-BA-057-A11 5', mRNAsequence

[a] only the most representative NCBI accession code is listed for each novel peptide; *Indicates the presence of a stop codon within the novel peptide sequence. Expect threshold was set to 1000 for all tblastn searches. Underlined peptide sequences represent tblastn hits with 100% identity and coverage.

Table S3 - Functional characterization and orthologous group identification of novel peptides identified by spectral correlations in EggNOG databases.

Novel peptide sequence	RNA accession number [a]	Predicted protein [b]	Function	Evalue	Score	Orthologous protein
ADKNLLR	XM_010055119.2	XP_010053421.1_1	Function unknown	4.49E-124	418.4	ENOG502QPRG
A*GESLCYGVPAR	XM_010026899.2	XP_010025201.1_1	Chromatin structure and dynamics	1.04E-75	259.1	KOG1756
KLG*LT*LR	XM_010037706.2	XP_010036008.1_1	Inorganic ion transport and metabolism	0	1,450.70	KOG0206
LSTMAAPTSGSGR	XM_010045996.2	XP_010044298.1_1	Function unknown	9.94E-181	604.3	ENOG502QR3U
NHGIHFR	XM_010034878.2	XP_010033180.1_1	Replication, recombination and repair	0	1,327.70	ENOG5037P6N
NLGSTANLR	XM_018860691.1	XP_018716236.1_1	Function unknown	0	1,345.40	ENOG502R4BG
NLGVGFAR	XM_010053962.2	XP_010052264.1_1	Transcription	4.05E-120	405.1	ENOG502QR2Y
REVCMGWHHFHAIGQR	XM_018864396.1	XP_018719941.1_1	Carbohydrate transport and metabolism	7.60E-187	626	ENOG501NSTY
SAENIMMPR	XM_010037390.2	XP_010035692.1_1	Function unknown	2.48E-116	391.3	NOG502QUIK
VALSDPIR	XM_018865801.1	XP_018721346.1_1	Function unknown	1.50E-135	455.4	ENOG502QSPM
VSLSIVPR	XM_010027007.2	XP_010025309.1_1	Function unknown	3.88E-242	807.7	ENOG502QQ35
RSGGEMTTAAPR [1]	XM_010068985.2	XP_010067287.1_1	RNA processing and modification	6.74E-196	654.7	ENOG5037I3W
RSGGEMTTAAPR [1]	XR_727223.2	-	RNA processing and modification	1.32E-155	521.9	ENOG5037I3W
EGLVSPLIK	XM_018869853.1	XP_018725398.1_1	Cytoskeleton	3.41E-116	381.1	ENOG5037SV9
SSTVKVTRK	XM_010062006.2	XP_010060308.1_1	RNA processing and modification	7.57E-86	269	KOG1847
VAVEVSPR	XM_018862919.1	XP_018718464.1_1	Cytoskeleton	4.32E-262	868	ENOG5037SSX
ARPFGAPR [2]	XM_018859771.1	XP_018715316.1_1	Secondary metabolites biosynthesis, transport, and catabolism	1.70E-123	415.9	ENOG5037NI4
ARPFGAPR [2]	XR_001979892.1	-	Secondary metabolites biosynthesis, transport, and catabolism	8.03E-13	52	ENOG5037NI4
RRNGVCLFYR [3]	XM_018861651.1	XP_018717196.1_1	Post-translational modification, protein turnover, and chaperones	6.82E-30	108.8	KOG3368
RRNGVCLFYR [3]	XR_001980976.1	-	Signal transduction mechanisms	1.26E-26	97.4	ENOG5037QKR
AASGNASLPR	XM_018861733.1	XP_018717278.1_1	Energy production and conversion	0	1,345.10	ENOG5037M9Y
GSGQTNCCSSRNPRNGR	XM_010046998.2	XP_010045300.1_1	Translation, ribosomal structure and biogenesis	7.29E-35	124.5	COG0359
HIQIGLR	XM_010039327.2	XP_010037629.1_1	Function unknown	0	1,487.60	ENOG5037HG6
KPEDANR	XM_010045581.2	XP_010043883.1_1	Function unknown	2.30E-271	906.2	ENOG5037IWW
LGTI**LSPR	XM_010065691.2	XP_010063993.1_1	Function unknown	5.20E-251	838.1	ENOG5037R5K
MNLPVKR	XM_018873041.1	XP_018728586.1_1	Function unknown	5.31E-77	263.5	ENOG502QQJE
RAIMRPR	XM_010029799.2	XP_010028101.1_1	Signal transduction mechanisms, Cytoskeleton	4.33E-277	925.8	ENOG5037IIP
KTAIEMSSMHEERK	XM_010054572.2	XP_010052874.1_1	Transcription	0	1,071.00	ENOG5037MN3
NLGLGFGR	XM_018876096.1	XP_018731641.1_1	Function unknown	3.89E-175	585.8	ENOG502QQ9S
APTTAPPAVR	XM_010062344.2	XP_010060646.1_1	Secondary metabolites biosynthesis, transport, and catabolism	1.04E-242	811	KOG1417
KIGLTLR	XM_010069929.1	XP_010068231.1_1	Function unknown	6.02E-47	164.1	ENOG501N7KN
KLV*ED*LAK [4]	XM_010046007.2	XP_010044309.1_1	Function unknown	2.70E-74	253.6	ENOG502RZV3
KLV*ED*LAK [4]	XM_018872208.1	XP_018727753.1_1	Carbohydrate transport and metabolism	1.04E-235	788.4	ENOG5037RH1
LFPLFSVR	XM_018870367.1	XP_018725912.1_1	Secondary metabolites biosynthesis, transport, and catabolism	9.52E-162	537.8	ENOG5037JSZ
LL*CC**QFPR	XM_018870037.1	XP_018725582.1_1	Translation, ribosomal structure and biogenesis, Transcription, Replication, recombination and repair	1.12E-195	647.6	KOG0335
QGKPTCPDHESTCPR	XM_010070621.1	XP_010068923.1_1	Inorganic ion transport and metabolism	9.15E-132	444.3	ENOG5037RFT
SVAIGERGLR	XM_010044558.2	XP_010042860.1_1	Function unknown	6.12E-157	526.4	ENOG502QTXK

TVVGLW*PCR	XM_018875477.1	XP_018731022.1_1	Transcription, Replication, recombination and repair, Signal transduction mechanisms	1.70E-66	229	ENOG502SHKV
IGTVTIPR	XM_018871868.1	XP_018727413.1_1	Function unknown	2.00E-130	438.5	ENOG502QRXU
PASSSRRPSR	XR_720815.2	-	No orthologs found	-	-	-
TVERLSMVEDVR	XR_001984062.1	-	Function unknown	2.01E-08	29.7	ENOG502SFCS
VTAAAADASPSR	XR_001979828.1	-	No orthologs found	-	-	-
APLALSPR	XR_721165.2	-	No orthologs found	-	-	-
EGDVLTLLESER	XR_001983327.1	-	Translation, ribosomal structure and biogenesis	2.58E-48	169.1	KOG3502
IPTGATLKGPMMTSPVLVEMAR	XR_001981593.1	-	Transcription	2.50E-07	34.4	KOG0627

[a] RefSeq database accession number in which the novel peptide was identified; [b] Predicted coding sequence for the NCBI mRNA sequences;
*Indicates the presence of a stop codon within the novel peptide sequence; No coding sequence was predicted for ncRNA and transcribed RNA databases as they were annotated as non-coding RNA sequences. Uppercase numbers indicate peptides with more than one mapping assignment.

CHAPTER 3 - Identification of novel peptides associated with seasonality and water restriction imposition in *Eucalyptus grandis* leaf proteome

ABSTRACT - Climate change is escalating the frequency and intensity of warming and drought periods around the globe, currently representing a threat to many plant species. Understanding how plants cope with such abiotic factors is crucial.We investigate how *Eucalyptus grandis*, a plant species with several industry applications, copes with seasonal variation and water restriction imposition at the proteomic level under field conditions. Therefore, we attempted to identify known proteins and novel peptides associated with the effect of seasonality and water restriction impositions, as well as to provide insights into how novel peptides behave under such conditions. Leaf proteome of *E. grandis* plants was studied under both a conventional proteomic workflow and a dedicated proteogenomics approach. The highest proteomic variability was identified in the summer season and the most abundant known proteins associated with seasonal variation were related to photosynthesis. Post-translational modifications, protein turnovers, and chaperones were the main functional classifications identified among biological treatments. Furthermore, 144 novel peptides were identified by both spectral correlations against modified databases (43) and a *de novo* peptide sequencing approach (101) not predicted by current proteomics pipelines. It is expected that most single amino acid substituted (SAS) peptides, mainly associated with the photosynthesis process, should decrease protein stability by altering the quantitative change upon ΔΔG values and non-covalent interactions. Multiple reaction monitoring validation assays were performed for selected novel peptide identifications demonstrating that it is a very robust mass spectrometry-based method of validation. Data are available via ProteomeXchange with identifier PXD031100.

Keywords: Proteomics, Proteogenomics, drought stress, MRM.

1. INTRODUCTION

Eucalyptus species are widely cultivated in many regions of the globe for their high adaptability to different climatic conditions and industry applications. Also, species from this genus show valuable adaptability to a wide range of habitats, which makes them one of the most planted trees in the world (Freeman *et al.*, 2013). Among them, *Eucalyptus grandis* species stands out for its versatility on paper and cellulose production, as well as in the oil industry and as charcoal. However, current environmental changes have negatively impacted yields and plant adaptability of this genus and many species of agricultural importance.

Global warming and climate change may exacerbate the effects of abiotic stresses on crop production as climate keeps changing (Li et al., 2020). As the world is running against time to propose and implement alternatives to cope with those effects, planted tree sector is a major partner in the fight to reduce the impacts of climate change through its forests and bio-based products (IBÁ - Indústria Brasileira de Árvores, 2020).

Global climate change is expected to increase the frequency and intensity of droughts occurring due to warming temperatures (Easterling et al., 2000). Drought, in particular, severely impairs plant growth and development and restricts crop productivity more than any other environmental factor (Shao *et al.*, 2009). Water stress alters cell structures and impairs various biological functions, thus leading to photosynthesis inhibition, metabolic dysfunction, membrane and protein damages (Krasensky and Jonak, 2012). In this context, proteins play a very important role within the plant adaptation to water stress since they regulate many cellular functions and are essential to the cellular machinery that enables different tasks within a given biological condition (Feist and Hummon, 2015). Thus, proteomics studies are crucial for understanding how plants behave under water stress conditions as they provide direct information on gene expression end products: proteins.

Protein identification in most proteomics workflows is a result of correlating experimental mass spectra data against known sequences in a reference protein database (virtual spectra). Such correlations provide solid evidence of the protein present in a particular biological sample (Eng *et al.*, 1994). Yet, proteogenomics, a field

of biological research that uses a combination of proteomics, genomics, and transcriptomics to aid in the discovery and identification of peptides, has emerged as an important strategy to search for novel peptide identities usually restricted by the conventional spectral correlation approach (Jaffe *et al.*, 2004). Additionally, novel peptide forms can be identified through the *de novo* peptide sequencing technique, whereas single amino acid substitutions (SAS) resulting from single nucleotide polymorphisms (SNPs) may also be detected (Nesvizhskii, 2014).

Analysis of shotgun proteomics data generally relies on the search against databases that do not include information on protein variants, usually preventing SASs from being identified (Li *et al.*, 2011). Consequently, single amino acid substitutions, which usually cannot be identified in reference protein database searches, are an important source of variability that may enhance plant adaptability and provide vital phenotype alterations. For instance, a single amino acid substitution (G41S) in a temperature-sensitive Fenoria allele has inhibited root hair growth in *Arabidopsis thaliana* plants (Kim *et al.*, 2021). Those SAS events can also have a high impact on protein thermodynamic stability, which can lead to protein structural changes (Kucukkal *et al.*, 2015).

Besides identifying novel peptides, targeted proteomics assays can provide an extra level of confidence for the existence of such peptides, and this technique has emerged as a powerful protein quantification tool in many biological systems (Shi *et al.*, 2016). Furthermore, it has been used for several applications such as investigating protein biomarkers in Loquat fruits for catalytic enzymes (Martínez-Márquez *et al.*, 2013), multiplexed MRM-based applications for biomarker discoveries, validation and accurate quantification in human disease studies (Percy *et al.*, 2013), and absolute quantitation of plant seed allergens (Houston *et al.*, 2011; Stevenson *et al.*, 2012).

As climate change has escalated seasonal changes and the occurrence of drought events, proteogenomics can play an important role in identifying novel targets that cannot be observed using conventional protein identification workflows and help to enhance our understanding on how plants cope with such climate changing. Therefore, the goal of this work is to confidently identify known proteins and novel peptides associated with the effect of seasonality and water restriction impositions, as well as to provide insights into how novel peptides behave under such conditions.

2. MATERIAL AND METHODS

2.1. Plant growth and experimental design

A commercial eucalyptus clone (clone number: O6) from the *Eucalyptus grandis* species was selected from an experimental site that belongs to TECHS Project (Tolerance of Eucalyptus Clones to Hydric, Thermal and Biotic Stresses) (Binkley *et al.*, 2017). The experiment was implemented on January 5th, 2017, at the Teaching, Research and Extension Farm (FEPE), Faculty of Agricultural and Veterinary Sciences, in Jaboticabal-SP, which is located at 21°14'11.2" south latitude and 48°16'53.7" west longitude with an average altitude of 604m. This clone has a plastic adaptation characteristic and an Af climate origin according to Köppen classification (Alvares *et al.*, 2013).

The experimental plot was formed by 5 plants in a linear distribution spaced 2m from each other and 3m between each row. Experimental plots were distributed under a complete randomized block design with two replications. Plants were grown under two contrasting water conditions: without water restriction (100% of rainfall) and with water restriction (approximately 30% of rainfall excluded). For the rainfall exclusion, a wooden platform was installed with inclined plastic channels between each row to carry part of the rainfall out of the experiment (Figure 1).

Figure 1 - Field experiment as part of TECHS Project located in Jaboticabal, Sao Paulo, Brazil. The image on the left illustrates Eucalyptus clones grown under 30% rainfall water restriction, while the image on the right shows the trees cultivated under normal rainfall conditions.

The experimental field is situated on a typical eutropheric Red Latosol, with a clay texture. After soil analysis, each plant was fertilized with 100g of ‘MG YORIN SI’

(YOORIN®). Cultural practices such as weeding, covering fertilization, and application of insecticides were carried out according to the plant's needs. Accumulated rainfall and temperature oscillation data was obtained by a meteorological station, located about 600m from the experimental site, and was recorded throughout the evaluation period (Figure 3).

2.2. Sample collection

Samplings were carried out over three seasons (winter: July 10th, 2019; spring: October 8th, 2019; and summer: January 08th, 2020). At the first sampling date, 31 months old trees were randomly selected, being the same plants studied in all sampling seasons. Young leaves from the upper-tree canopy were individually collected from each tree under both experimental conditions: with and without water restriction. They were immediately frozen in liquid nitrogen and stored at -80◦C before protein extraction.

2.3. Protein extraction and digestion

Two grams of leaf material from each treatment were individually processed, macerated in liquid nitrogen until a fine powder was achieved. Each sample was extracted during 15 min (at 4°C) through addition of 12 mL Tris-HCl buffer solution (125 mM Tris pH 6.8-HCl, 20% (v/v) glycerol, 1% (w/v) SDS, 1% (w/v) DTT) and stayed overnight precipitating in acetone (-30°C) (sample/acetone ratio of 1:5). The protein pellet was then resuspended in a 150 µL buffer solution, and protein concentration was measured using Bio-Rad Protein Assay kit according to manufacturer's instructions. Before mass spectrometry analysis, a 50-µg aliquot of leaf proteins was loaded into sodium dodecyl sulfate-polyacrylamide electrophoresis (SDS-PAGE) gel and shortly separated for 1 hour. Then, the resulting 2 cm-wide bands were sliced into smaller gel pieces and transferred to clean tubes to be reduced, alkylated, and *in-gel* digested (Shevchenko *et al.* 2007). Subsequently, the digested peptides were submitted to two

filtering steps using a Costar® Spin-X® centrifuge tube filter (Salt Lake City, Utah, USA) and a Millipore C18 Ziptip® (Billerica, MA, USA).

2.4. Acquisition of mass spectrometric data

After filtration step, three replicates of each sampling season under both experimental conditions (with and without water restriction) were individually resuspended in 20 µL (0.1% (v/v) formic acid), totalizing 18 samples. An aliquot of 10 uL of each replicate was randomly injected under a chromatographic separation of 65 min in an Easy nLC 1000, Thermo Scientific® coupled to an Exactive™ (Thermo Scientific) system operated in data-dependent acquisition (DDA) mode for the 20 most abundant peptide ions. Peptide separation was achieved with a binary buffer system consisting of 0.1% formic acid (buffer A) and 99% acetonitrile in 0.1% formic acid (buffer B), using a linear gradient of buffer B (from 7% to 30% in 50 min, and at 70% for other 15 min). Each sample injection was followed by two washouts (constant at 95% of B for 20 min, ramping down to 5% of B for 5 min) at a flow rate of 400nl/min. Acquisition of high-resolution mass spectra was carried out concomitantly to the chromatographic separation. Full (400-1600 m/z) and data-dependent scans were acquired with resolutions equal to 70,000 and 17,500 FWHM (full width at half maximum), respectively. Fragment spectra were acquired upon HCD (high-energy collision dissociation) fragmentation of the isolated peptide ions under normalized collision energy of 30 eV. Peptide ions selected for fragmentation were excluded in a new isolation event for 5 s.

2.5. Known proteins and novel peptides' identification by spectral correlation

Generated spectra of all raw files were analyzed using MaxQuant software (v 1.6.14.0) (Cox and Mann, 2008; Tyanova *et al.*, 2016a) by means of Andromeda search engine (Cox *et al.*, 2011). For known protein identifications, peak lists were searched against the *E. grandis* v. 2.0 protein database (46,280 entries), available for download at Phytozome database (Goodstein *et al*., 2012). However, the database used for novel peptide identifications by spectral correlations was created from mRNA, ncRNA, and transcribed RNA RefSeq databases as described by Jorge and Balbuena (2021). The searches were performed with most of the default settings with some adjustments as follows: specific digestion mode was set to trypsin (P) and a maximum of two missed cleavage sites was allowed; carbamidomethylation (C) as fixed and oxidation (M) as variable modifications; minimal peptide length was defined as 7 amino acids; label-free relative quantitation algorithm (MaxLFQ) (Cox *et al.*, 2014) was used for comparative analysis between different treatments, and LFQ minimum ration count was set to 1. Using a target-decoy approach, false discovery rate (FDR) at protein and PSM levels was set to 1%. Normalized LFQ intensities were used for further data analysis. Mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via PRIDE partner repository (Perez-Riverol *et al.*, 2022) with dataset identifier PXD031100.

2.6. Novel peptide identifications by *de novo* sequencing

PEAKS Studio 10.5 (Bioinformatics Solutions, Waterloo, Ontario, Canada) software was used to provide automated *de novo* sequencing from MS/MS spectra. This approach involved three major steps such as PEAKS *de novo* sequencing, followed by a PEAKS DB search (Zhang *et al.*, 2012) against the *E. grandis* v.2.0 protein database, and finally, was used SPIDER (Han et al., 2005) algorithm integrated within the PEAKS software. Detailed *de novo* sequencing workflow and criteria used for the identification of novel peptides with single amino acid substitution (SAS) were performed according to Jorge and Balbuena (2021). The area under the curve quantitative value of each significant SAS peptide was selected for further analyses.

2.7. Statistical analysis

Perseus (v 1.6.14.0) was used for data organization and statistical analyses of all identifications (Tyanova *et al.*, 2016b), including results from both spectral correlation and *de novo* identifications. Protein identifications that matched the reverse database, contaminants, and proteins identified only by modified amino acids were eliminated. Each known protein or novel peptide identification was firstly classified according to its biological replicates and experimental conditions, and data were log2 transformed. Before each specific comparison, only known proteins or novel peptides quantified in three biological replicates for at least one experimental condition were taken into account for statistical analyses. Missing values were imputed from a normal distribution of each biological sample with the default settings (width: 0.3; downshift: 1.8).

Pairwise comparison analyses were carried out separately for each specific season to compare plants under both conditions: with and without water restriction. Student's t-test was performed to identify significant changes in abundance between both conditions. Identifications with a p-value <0.05 and, at the same time, an absolute fold-change value greater than 2 were defined as significantly down or up-regulated peptides/proteins. However, regarding comparisons across different seasons (winter, spring, and summer), an ANOVA test was performed with a permutation-based FDR of 0.05 for the identification of significant abundance changes.

2.8. Data reasoning and visualization

A principal component analysis (PCA) was performed comparing the ANOVA significant known proteins among all six conditions. Also, hierarchical clustering (Euclidean distance and average linkage) was performed for ANOVA significant identifications using Euclidean distance and Perseus default program parameters.

To enable an enrichment analysis of the most abundant known proteins across different seasons, BLASTP tool of Blast2GO software was used against revised sequences data from *A. thaliana*, obtaining the accessions of the corresponding proteins in *A. thaliana* for the identifications of the current work. Thus, functional

annotation information available for *A. thaliana* were inferred for the proteins identified here. Gene Ontology analysis (GO annotation) and KEGG Pathway identified for each high abundant known protein of interest were subjected to enrichment analysis by Fisher's exact test in Perseus software. Thus, it was possible to identify which GO categories (Ashburner *et al*. 2000) and KEGG Pathway (Kanehisa and Goto, 2000) were enriched among such proteins. All eligible known protein identifications (after passing through the minimum number of valid values filter) were used as a background cluster. A Benjamini-Hochberg FDR was used for false discovery correction rate and a threshold value was set to 0.02.

The overall taxonomic assignment and functional annotation of known proteins and novel peptides were performed following eggNOG-mapper v2 (Cantalapiedra *et al*. 2021). The highest-scoring annotated Cluster of Orthologous Group (COG) (Tatusov *et al.* 2003) category hit was selected for each functional annotation. For novel peptide identifications achieved by spectral correlations, the predicted protein-coding sequence for each NCBI RefSeq RNA accession number was considered for the reference annotation.

Genomic mapping of novel peptides identified by spectral correlations against modified RNA databases was carried out by TBLASTN searches against the Annotation Release 101 from *E. grandis* v.1.0 database and the usage of the NCBI Genome Data Viewer (GDV), which is a tool embedded in the TBLASTN result page. Also, the visualization of known and novel identification positions at the chromosome level was accomplished by the ChrDiagram tool, publicly available at https://eucgenie.org/. As the tool requires accession numbers from the *E. grandis* v.2.0 protein annotation, a protein-protein basic local alignment search tool (BLASTP) was carried out for all novel peptides found by spectral correlations as they were initially identified in the *E. grandis* v.1.0 database (Supplementary Table 1).

2.9. Single amino acid substitution structural analysis

First of all, for each t-test or ANOVA-significant novel peptides identified by *de novo* sequencing approach, a protein template structure was searched in a non-redundant PDB library by the sequence search using MMseqs2 method (Steinegger and Söding, 2017) to identify similar protein chains in PDB Archive (Berman *et al.*, 2000). Protein model selection was based on the following criteria: a) presence of the same native amino acid between both protein sequences; b) higher similarity between the whole peptide sequence within both proteins; and c) minimum of 50% sequence identity to the selected protein template and the native protein sequence. The complete description of evaluated parameters for protein structural analyses, as well as protein template selection details, can be found in Supplementary Table 9.

After selecting the appropriate templates, computational tools (structure-based) PremPS (Chen *et al.*, 2020) and mCSM (Pires *et al.*, 2014) were used to calculate the influence of single amino acid substitutions on the ΔΔG value of each protein. 3D virtual native and substituted protein structures were displayed by PremPS to illustrate non-covalent relationships between the substituted site and its neighboring residues. As PDB structure templates were not identified for all single amino acid substituted candidates, predictions of Gibbs free energy were carried out by a sequence-based approach, Mupro (Cheng *et al.*, 2006). These tools are available as web services and allow submission and fast calculation of large sets of single amino acid substitutions.

2.10. Validation of novel peptides and single amino acid substitutions

2.10.1. Synthetic peptide library

Out of 144 novel peptides identified through proteogenomics approach, 24 were selected for the validation experiment based on the following criteria: novel peptides identified by spectral correlations were selected according to their different classes of genomic classifications and higher abundance level. Furthermore, the selection of candidate peptides identified by *de novo* peptide sequencing approach was performed based on peptide uniqueness, presence of only one peptide per accession code (protein), internal peptides. Similar peptide sequences were excluded from the

analysis, novel peptides were identified only in the substituted form, and both forms (native and substituted) and higher intensity and fold change between substituted and native peptides were prioritized (Supplementary Table 10). 24 PEPscreen® peptides (Custom Peptide Libraries, Sigma-Aldrich, St. Louis, USA) were synthesized with a yield of ~ 1–6 mg for the validation experiments. Before biological testing, PEPscreen® peptides were diluted in 50% ACN (5 mM stock), and stored at − 80 °C.

2.10.2. MRM method optimization

Preliminary transitions were firstly optimized according to Skyline (64bit V21.1.0.278) MRM method (MacLean *et al.*, 2010; Bereman *et al.*, 2012). Precursor charges +2 and +3, fragment charges +1 and +2, as well as y and b ion types, were considered for all synthetic peptides according to each amino acid sequence. Before mass spectrometry injections, all 24 synthetic peptides were combined into one pool at an equimolar concentration of 100 femtomoles μL^{-1}. Mass spectra were individually acquired for each peptide over a gradient of 30 min of acquisition. Subsequently, a minimum of 4 and a maximum of 6 most intense peaks in the MS/MS spectrum were manually selected as final MRM transitions. Additional runs were carried out for three peptides at a concentration of 200 femtomoles μL^{-1} due to weak fragment ion intensities. After selecting the best candidate transitions for each synthetic peptide, the optimized collision energy for each transition was tested. For each peptide, collision energy optimization parameters were set to use five steps on either side of the value predicted by the default equation, with the step size set to 1 V. Therefore, there were 11 collision energy voltage values for each transition. Another round of mass spectra was acquired for each peptide and the best collision energy. The one that generated the highest ion peak area intensity response was selected for each transition. Optimized transition and collision energy values were imported into a final, optimized MRM method (Supplementary Table 11). Lastly, the linear range response for each peptide was defined by 7 concentrations at 1:4 dilution ratios (0.064 femtomolaes μL^{-1} to 1 picomoles μL^{-1}) (Figure S4). Three technical replicates were acquired at each concentration for each peptide (a total of 136 transitions were monitored for all 24 synthetic peptides at each injection).

2.10.3. Mass spectrometry data acquisition *(LC-MRM)*

Peptides (5 µL) were loaded onto a C8 trap column (Thermo Scientific 161194, u-Precolumn, C8 PepMap100, 5 mm x 300 µm, 5 µm particles) through a 3 min gradient at 5µL min^{-1}, and separated on a 20 cm x 75 µm self-packed fused silica column containing HxSIL C18 (5um particles, The Hamilton Co.) An Eksigent nanoLC Ultra 1D Plus was used for the reverse-phase liquid chromatography over a 30 min gradient at 600 nL min^{-1} as follows: initial conditions were 2% B (A: 0.1% formic acid in water; B: 0.1% formic acid in acetonitrile); a 15-min linear gradient from 2 to 90% solution B, followed by a linear gradient from 90 to 40% solution B (3 min); oscillating column wash: 80% B (2 min), 40% B (2 min), 80% B (2 min), 2% B (1 min); and held at initial conditions for 5 min. Acquisition of mass spectra was carried out concomitantly to the chromatographic separation. Peptides were analyzed through a Multiple Reaction Monitoring (MRM) method using a Quantiva QQQ mass spectrometer (Thermo Scientific). The instrument was operated in nano Electrospray positive-ion in un-scheduled MRM mode with the following parameters: spray voltage of 1800 V; 1 min cycle time; Q1 filter=1.2 Da; Q3 filter=0.7 Da; and CID gas = 3 mTorr.

2.10.4. Validation criteria for novel peptides

Protein extraction and digestion were carried out following the same protocol as described in section 2.3. After the filtration step, three biological replicates of each condition were combined into a pooled sample of 20 µL volume (0.1% formic acid; 5% ACN), totalizing 6 pool samples. An aliquot of 5 µL of each pooled sample was randomly injected following the same acquisition method described in the section 2.10.3. For the retention time calibration, all conditions were combined into one pool sample and all synthetic peptides were spiked in the biological sample to a final concentration of 40 femtomolar $µL^{-1}$ to account for any biological matrix effects and retention time shift that could occur during mass spectrometry analysis. Aligned transitions were manually checked for each peptide on Skyline and the following criteria for novel peptides validation was considered: a) minimum of 3 aligned transitions for each peptide (retention time shift < 0.2 minutes between transitions); b)

intensity of each identified transition for the biological samples higher than the intensity of the synthetic peptide injection at a minimum valid concentration. Such valid concentration was visually determined based on the lowest concentration at which each synthetic peptide showed a well-defined alignment of all transitions; c) retention time match between the biological sample injection and the spiked-in sample (maximum shift of 0.2 minutes); d) similar intensity proportions between synthetic peptide and biological sample injections for the selected transitions, and e) biological samples analyzed within each synthetic peptide's calibration curve previously determined (Figure S4).

3. RESULTS

3.1. Identification and characterization of genome-predicted proteins and novel peptides

Global proteome analysis revealed the identification of genome-predicted proteins and novel peptides mapped to different chromosome locations according to the *Eucalyptus grandis* BRASUZ1 genome annotation v2.0 (Figure 2). Mass spectrometric data acquisition and spectral correlation searches against *E. grandis* v. 2.0 protein database allowed the identification of 1031 known proteins (Supplementary Table 2); while proteogenomics approach used in this study revealed the identification of 144 novel peptides that had not been predicted by current gene annotation methods. They were identified following a dedicated proteogenomics workflow based on both spectral correlations against modified Refseq RNA databases (43) and *de novo* peptide sequencing (101) (Supplementary Table 3).

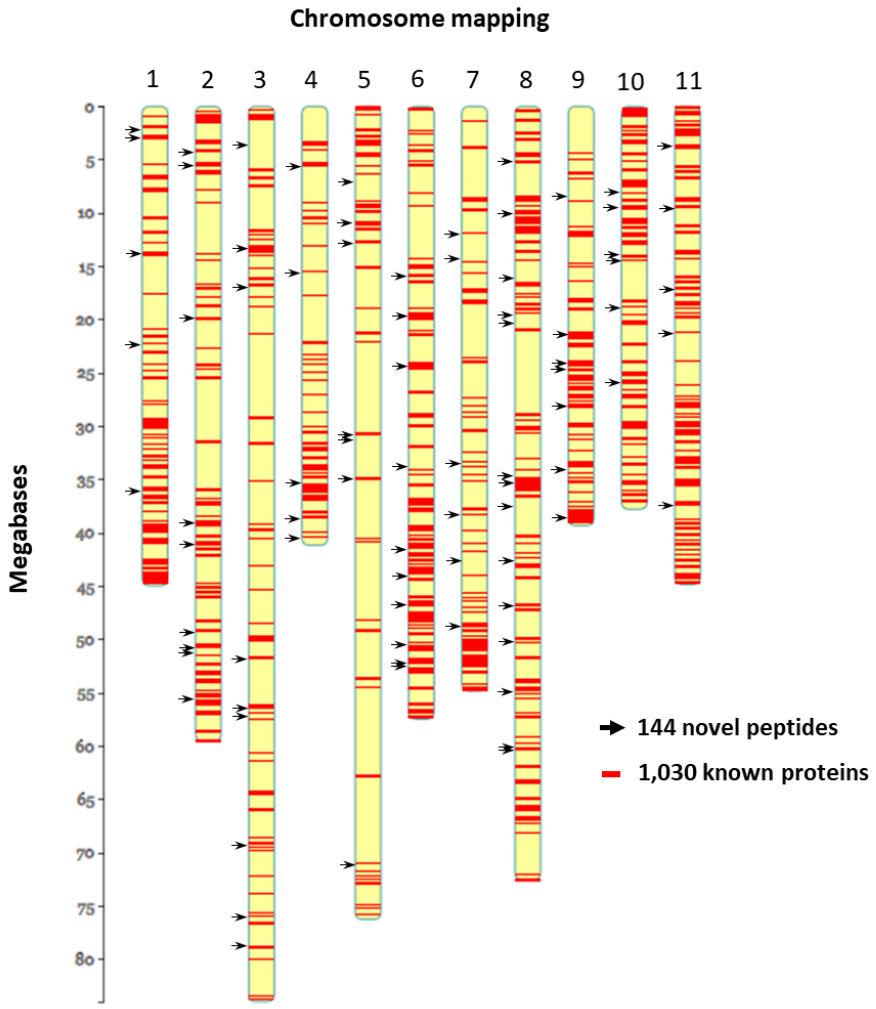


Figure 2 – Genome-predicted protein and novel peptide chromosome mappings according to the *Eucalyptus grandis* BRASUZ1 genome annotation v2.0.

Novel peptides identified by spectral correlations against modified RNA databases were further categorized regarding their genomic location (Supplementary Table 4). Genomic mapping of the 43 novel peptides indicated the existence of previously unknown coding sequences within *E. grandis* proteome, which allowed 43 unique mapping assignments. Depending on the genomic location of the new coding sequences, novel peptides were classified into the following categories: 5' UTR, 5' UTR-EXON, EXON-ARF, 3' UTR, and Novel Coding Locus (Figure S1).

3.2. Seasonal-based changes and their impact on protein variability

Weather parameters, including precipitation and temperature levels, consistently varied across different sampling periods (Figure 3, panel A). Precipitation levels were close to zero millimeters in the winter season, and reached the highest values in the summer. Furthermore, a Principal Component Analysis (PCA) generated six distinctive groups corresponding to each biological condition (Figure 3, panel B). Only known proteins with significantly different LFQ intensities (ANOVA, p-value ≤ 0.05) between all six experimental conditions were considered in this analysis.

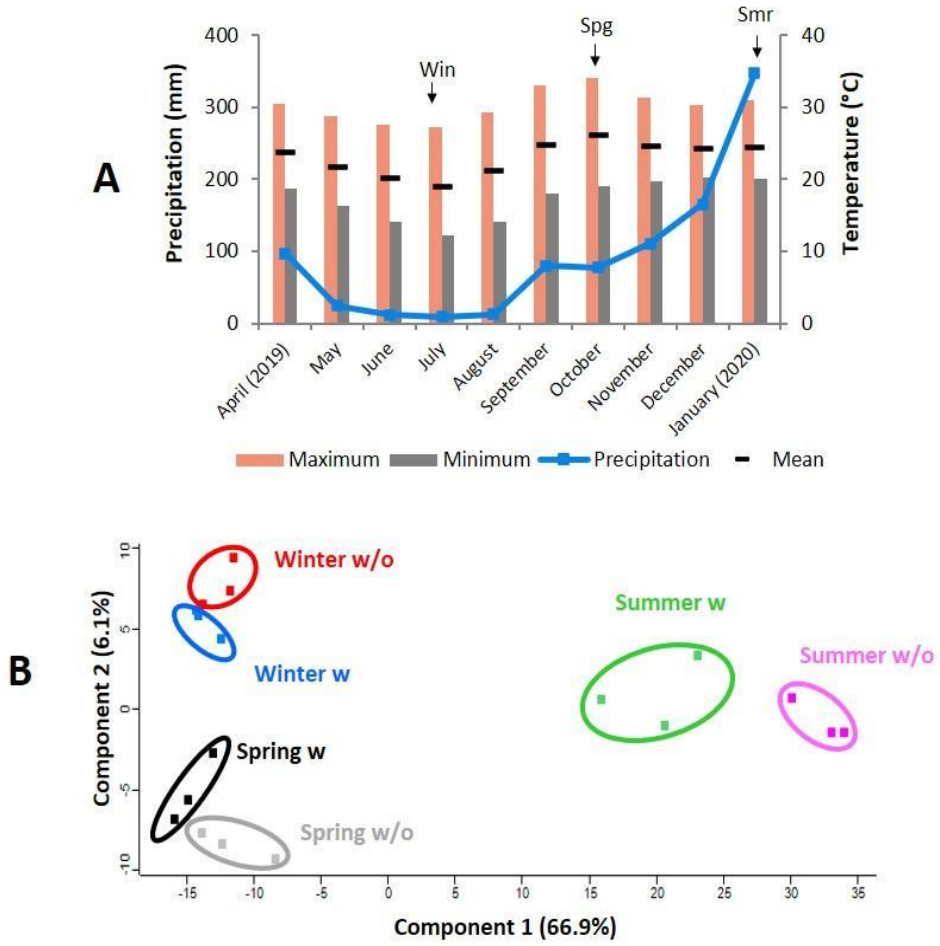


Figure 3 – Climatic parameters and Principal Component Analysis (PCA) of *Eucalyptus grandis* leaf protein samples. (A) Accumulated precipitation (blue line) and temperature oscillations (colored bars) at the TECHS site located in Jaboticabal, Sao Paulo, Brazil. Season abbreviations indicate the months in which samples were collected. (B) PCA analysis based on LFQ intensities obtained from ANOVA-significant known proteins among all treatments. The following symbols represent w/o: without water restriction; w: with water restriction.

Differences in proteome profiles were predominant between all biological conditions although samples collected during the winter were more closely grouped than samples collected in spring and summer. Data variability of *Eucalyptus grandis* leaf proteome was more pronounced towards different growing seasons rather than towards water restriction. Also, when comparing samples collected during the same season, water restriction seemed to have a higher impact on proteome variability in the summer.

3.3. Most responsive proteins to seasonal variation and their functional classification

Hierarchical clustering based on LFQ intensities of ANOVA significant known proteins across different seasons is shown in Figure 4. The accession numbers of the most abundant *E. grandis* protein groups and common identifications between different seasons of the year with and without water restriction are also presented. 280 proteins significantly varied under water-restricted condition and were visually grouped into eight different clusters; while 360 proteins significantly varied under no water-restricted condition and those proteins (visually grouped into seven clusters). A similar pattern of protein abundance could have been identified when analyzing both winter and spring seasons under each water regime.

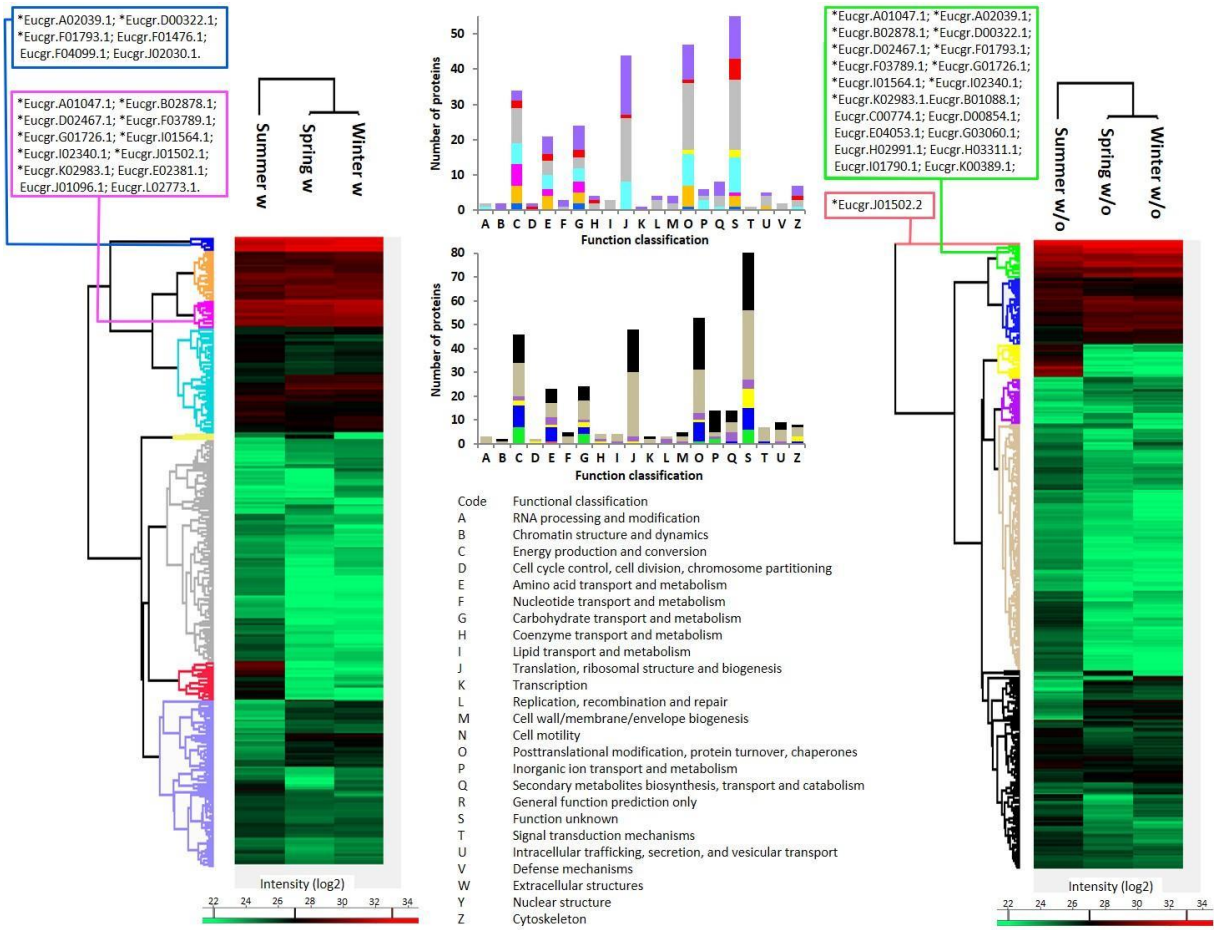


Figure 4 - Hierarchical clustering visualization analysis based on ANOVA-significant known proteins and functional classification according to COG categories. MaxLFQ intensity of each protein identification is displayed as log2 transformed data under different color codes (red: most abundant; green: least abundant). For the hierarchical clustering, the following symbols represent w/o: without water restriction; w: with water restriction; * indicates common proteins identified across both comparisons. In the bar graphs, capital letters on the x-axis indicate the cluster of orthologous groups (COG) and the y-axis indicates the number of known proteins identified in each class. Different color codes for the visually grouped clades from the horizontal hierarchical clustering are matched to the corresponding functional classification in bar graphs.

Proteins with the highest abundance levels were highlighted in two clades for water restricted and non-restricted conditions (blue and pink colors, and orange and green colors, respectively). The most abundant proteins among those clades were present in functional classes such as energy production and conversion, amino acid transport and metabolism, carbohydrate transport and metabolism, post translational modification, protein turnover, chaperones, and function unknown. Also, this functional classification pattern was observed among other known protein groups under both

water regimes at lower abundance levels. Furthermore, functional classification: translation, ribosomal structure and biogenesis had also a high number of protein identifications at intermediate levels of abundance. Yet, the unknown functional classification revealed the highest number of protein identifications throughout the evaluated conditions.

The most abundant known proteins identified in *E. grandi*s leaf samples over different seasons (across both water restricted and non-restricted conditions) was also compared based on an enrichment analysis. Such proteins were distributed under several categories of enriched functional classifications under either cellular component, biological process, molecular function, or KEGG pathway, and were mainly associated to the photosynthesis process (Supplementary Table 5). Enriched categories had enrichment factor values greater than 1, showing that these GO categories are enriched among the most abundant proteins across different seasons regardless of water restriction imposition. Enrichment profile in a highly abundant cluster of known proteins in *E. grandis* leaves reinforces the crucial role of photosynthesis for plants. Nevertheless, under no water restriction imposition, only biological processes and KEGG pathway were identified as enriched terms, and showed a lower enrichment factor when compared to the water restricted condition treatment.

3.4. Effect of water restriction on known proteome variability

Higher proteomic variability was observed for *E. grandis* trees under water restriction during the summer season, in which water restriction imposition down-regulated the abundance of 50 protein groups (Figure 5).

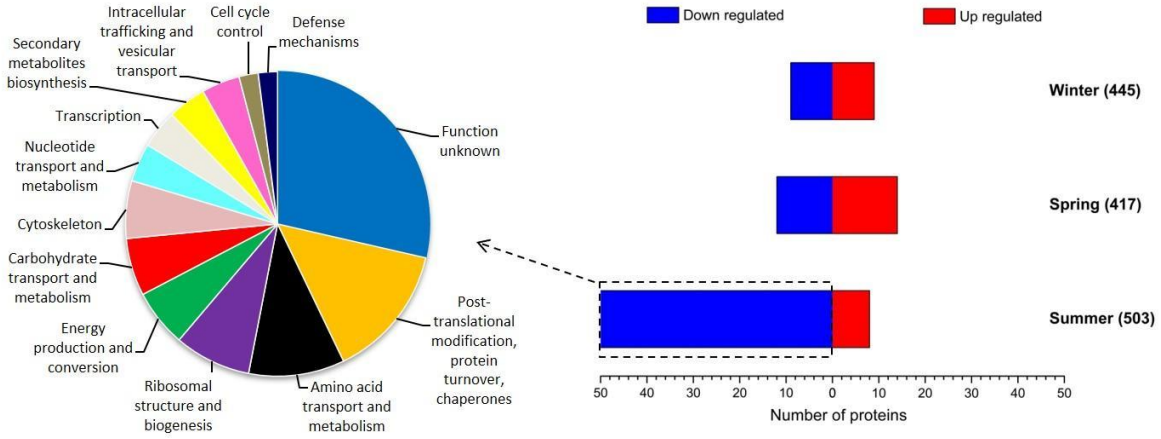


Figure 5 - Number of differentially abundant proteins based on pairwise comparisons of water-restricted against non-water-restricted *E. grandis* leaf samples for each season, and functional classification of down-regulated proteins identified in the summer. Numbers in parentheses represent quantified proteins eligible for pairwise comparisons after the minimum number of valid values filtration step. Only identifications with a significant Student's t-test (p-value < 0.05) and fold-change ≤ or ≥ 2 values were considered as a significant down or up-regulated protein, respectively. Pie chart indicates functional classification according to COG categories. Detailed functional analysis of all significant identifications can be found in Supplementary Table 6.

A significant part of the down-regulated proteins in the summer was classified into molecular functions such as post-translational modification, protein turnover, and chaperones, as well as amino acid transport and metabolism. On the other hand, nine protein groups were up-regulated under water restriction in the same season. A more proportional distribution between up and down-regulated protein groups was observed in winter and spring. Winter revealed the lowest number of differentially abundant proteins, which were nine up and nine down-regulated proteins.

3.5. Novel peptides associated with seasonal variation

Out of the 101 novel peptides identified by a *de novo* peptide sequencing approach, eight novel peptides showed a significant abundance variation across different seasons without water restriction (Figure 6, panel A). On the other hand, 10 out of 44 novel identifications achieved by spectral correlations against modified RNA

Refseq databases demonstrated a significant abundance variation among the different seasons under no water restriction (panel B).

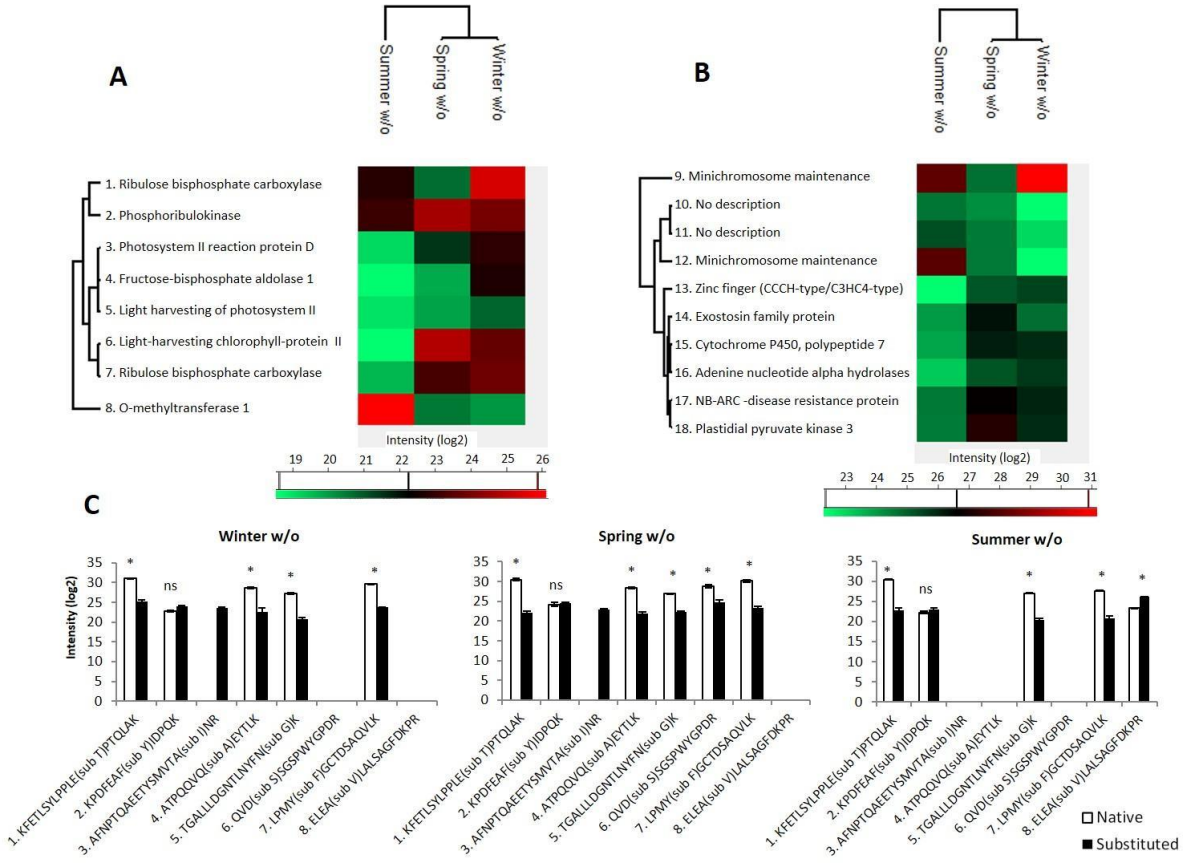


Figure 6 – Novel peptides' variability according to seasonal changes in non-water-restricted *E. grandis* leaf samples. The hierarchical clustering visualization analysis was based on ANOVA-significant novel peptides identified by (A) *de novo* peptide sequencing and (B) spectral correlation approaches. Quantitative values of novel peptide identifications are displayed as log2 transformed data. For the hierarchical clustering, the following symbols represent w/o: without water restriction; w: with water restriction. (C) Significant novel peptide identifications by *de novo* peptide sequencing were compared based on their native vs. substituted quantitative abundance levels according to Student's t-test (p-value < 0.05). Values are means ± SE of 3 replicates. * Indicates a significant protein abundance change. Missing values are displayed as gaps in the bar graphs. The detailed description of significant novel peptide identifications can be found in Supplementary Table 7.

Across different seasons under water restriction, novel identifications showed inferior variability levels. Three novel peptides identified by spectral correlations revealed a significant abundance variation, while only one novel peptide identified by the *de novo* peptide sequencing approach presented a significant variation among

seasons (data not shown). Additionally, regarding novel peptides identified by *de novo* sequencing, most native peptides revealed higher abundance levels when compared to substituted peptides (panel C). In contrast, a small number of novel peptides indicated non-significant abundant changes or were not present in the compared conditions. Only peptide 8 - ELEA(sub V)LALSAGFDKPR (O-methyltransferase 1) showed a higher abundance level for the substituted form when compared to its native form.

3.6. Novel peptides associated with water restriction imposition

Novel peptides identified through *de novo* peptide sequencing approach were predominant among peptides significantly associated with water restriction imposition in each specific season (Figure 7, panel A). In one case the same peptide was present in winter and summer (YDGPNSTDFIW(sub E)DWVK). In both cases, their abundance levels were higher under water restriction imposition. Alternatively, novel peptides identified by spectral correlations pointed out a less predominant association with water restriction imposition as only two peptides were significantly different (one in the spring and another in the summer).

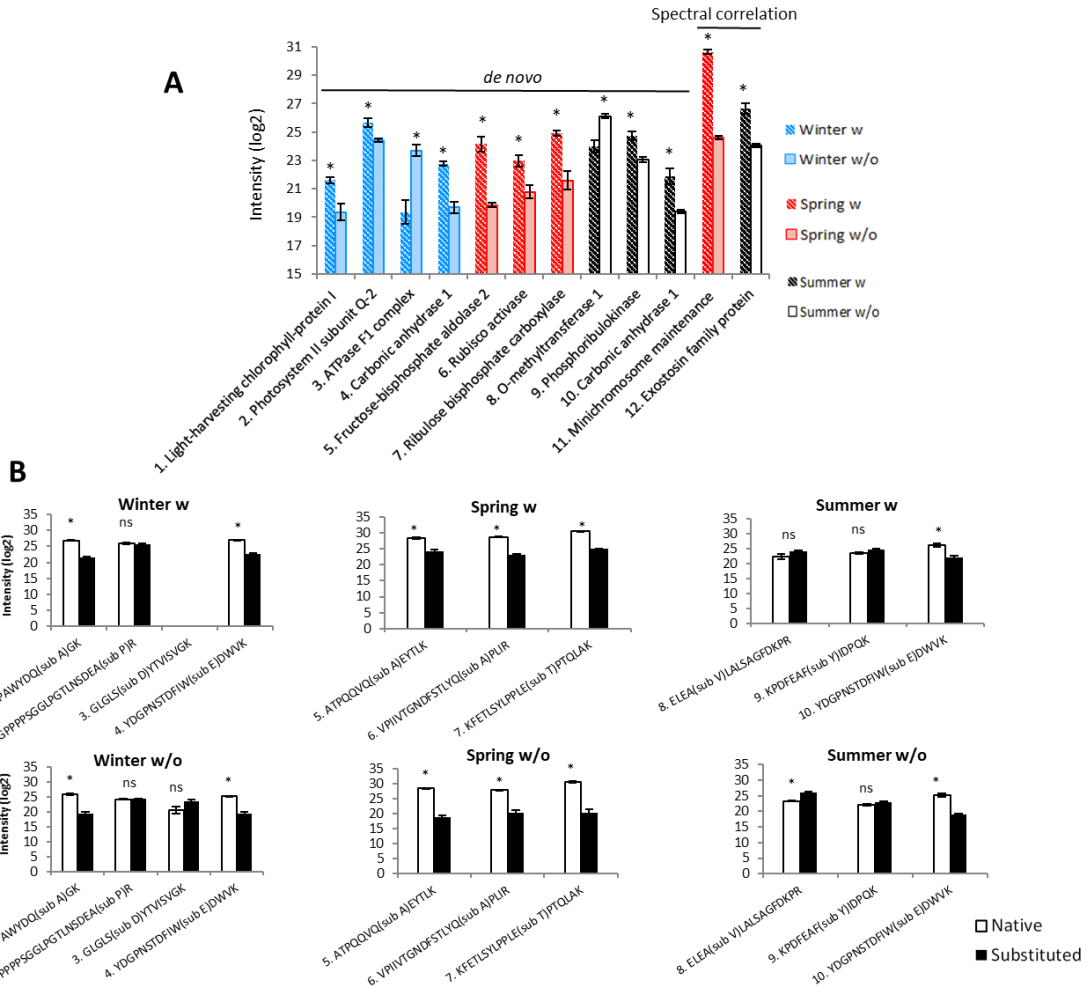


Figure 7 – Novel peptides' variability of *E. grandis* leaf samples according to water restriction imposition during each season. (A) Differentially abundant peptides (identified by both spectral correlation and *de novo* sequencing) based on pairwise comparisons of water-restricted against non-water-restricted *E. grandis* leaf samples in each season. Quantitative values of novel peptide identifications are displayed as log2 transformed data. Only identifications with a significant Student's t-test (p-value < 0.05) and fold-change ≤ or ≥ 2 values were considered as a significant down or up-regulated protein, respectively. (B) Significant novel peptide identifications by *de novo* peptide sequencing were compared based on their native vs. substituted quantitative abundance levels according to Student's t-test (p-value < 0.05). Values are means ± SE of 3 replicates. * Indicates a significant protein abundance change. Missing values are displayed as gaps in bar graphs. The detailed description of significant novel peptide identifications can be found in Supplementary Table 8.

Out of the 12 novel peptide identifications in both *de novo* and spectral correlation approaches, 10 significantly demonstrated superior abundance levels when *E. grandis* trees were under water restriction. Pairwise comparisons of the novel peptides identified by *de novo* sequencing also indicated a superior abundance

accumulation of native peptides when compared to substituted peptides (panel B). In some cases, native and substituted peptides did not significantly vary when contrasting their abundance levels at each specific biological condition, except for peptide ELEA(sub V)LALSAGFDKPR (O-methyltransferase 1) that showed a higher abundance of its substituted form in the summer without water restriction.

3.7. *De novo* peptide structural analysis

The structural analyses of significant novel SAS peptides that showed different abundance levels across different seasons under no water restriction imposition (Figure 7, panel A) are presented in Table 1. Significant identifications in each season with water restriction imposition (Figure 8, panel A) are shown in Table 2. Computational approaches PremPS, mCSM (structure-based), and MUpro (sequence-based) were used to calculate the influence of single amino acid substitutions on the ΔΔG value. PremPS and mCSM stability analyses were based on the most similar protein sequence structure available in the PDB repository, while for MUpro tool, the full native protein sequence was used for each SAS peptide.

TABLE 1 - Protein complex stability upon single amino acid substitutions on ANOVA-significant peptides.

Peptide sequence	Position	PremPS[a]	mCSM[b]	Position	MUpro[b]
1.KFETLSYLPPLE(sub T)PTQLAK	T22E	0.97	0.21	T80E	-0.42
2. KPDFEAF(sub Y)IDPQK	Y185F	0.09	-0.89	Y209F	-0.61
3. AFNPTQAEETYSMVTA(sub I)NR	-	-	-	I193A	-1.79
4. ATPQQVQ(sub A)EYTLK	A244Q	1.65	-1.81	A197Q	-0.45
5. TGALLLDGNTLNYFN(sub G)K	G114N	0.66	-0.88	G166N	-1.62
6. QVD(sub S)SGSPWYGPDR	S11D	0.25	-0.28	S44D	-0.44
7. LPMY(sub F)GCTDSAQVLK	F75Y	0.36	-0.44	F131Y	-1.44
8. ELEA(sub V)LALSAGFDKPR	-	-	-	V329A	-1.03

[a] Values of ΔΔG (kcal/mol)>0 mean destabilization; [b] Values of ΔΔG (kcal/mol) < 0 mean destabilization.

The vast majority of single amino acid substitutions indicated a loss in protein stability except for only one SAS peptide for the mCSM method. Negative values of ΔΔG for mCSM and Mupro indicate a loss in protein stability, while positive values of ΔΔG for PremPS suggest protein stability loss. Significant identifications of both t-test and ANOVA comparisons showed a similar pattern of protein stability decrease on

SAS novel peptides. Detailed parameters for protein structural analyses and protein template selection on both t-test and ANOVA-significant novel peptides are shown in Supplementary Table 9.

TABLE 2 - Protein complex stability upon single amino acid substitutions on t-test-significant peptides.

Peptide sequence	Position	PremPS[a]	mCSM[b]	Position	MUpro[b]
1. IGIINVPAWYDQ(sub A)GK	A129Q	0.47	-0.54	A171Q	-0.68
2. GPPPPSGGLPGTLNSDEA(sub P)R	-	-	-	P109A	-1.81
3. GLGLS(sub D)YTVISVGK	-	-	-	D163S	-1.85
4. YDGPNSTDFIW(sub E)DWVK	E244W	0.47	-0.32	E294W	-1.11
5. ATPQQVQ(sub A)EYTLK	A244Q	1.65	-1.81	A197Q	-0.45
6.VPIIVTGNDFSTLYQ(sub A)PLIR	A234Q	1.57	-1.16	A293Q	-0.98
7.KFETLSYLPPLE(sub T)PTQLAK	T22E	0.97	0.21	T80E	-0.42
8.ELEA(sub V)LALSAGFDKPR	-	-	-	V329A	-1.03
9. KPDFEAF(sub Y)IDPQK	Y185F	0.09	-0.89	Y209F	-0.61
10. YDGPNSTDFIW(sub E)DWVK	E244W	0.47	-0.32	E294W	-1.11

[a] Values of ΔΔG (kcal/mol)>0 mean destabilization; [b] Values of ΔΔG (kcal/mol) < 0 mean destabilization.

For a selected protein example (Figure 8), a 3D virtual display from PremPS shows the non-covalent relationships between the substituted site and its neighboring residues, as well as the native site and its neighboring residues. A threonine (native form) to glutamic acid (substituted form) substitution at position 22 of 5IU0 PDB template resulted in a loss in some hydrogen bonds, as well as a shift from Hydrophobic to Van der Waals interaction. Thus, those modifications within the protein non-covalent bonds suggest a loss in protein stability. The non-covalent interactions and illustration details for the other significant SAS peptides are shown in Figure S2.

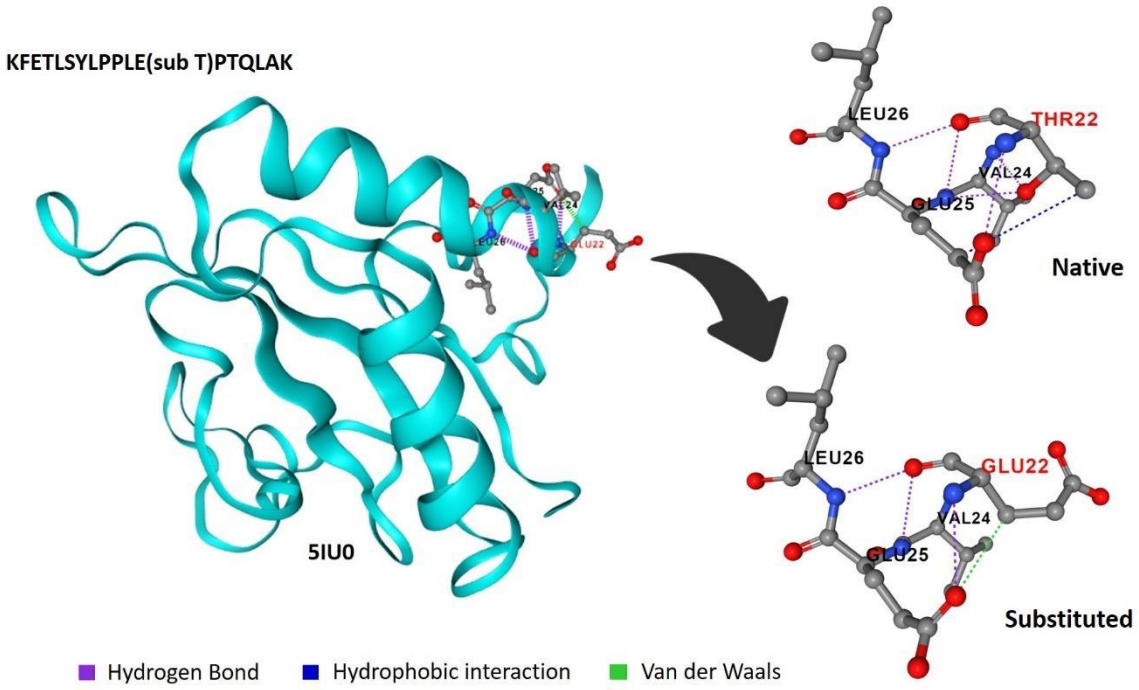


Figure 8 – Comparison of non-covalent interactions formed by a single amino acid substitution residue for the native and substituted protein forms in a selected example. Novel peptide sequence is shown above each reference PDB protein structure.

3.8. Novel peptides validation assays

Multiple reaction monitoring (MRM) assays were performed to validate the existence of the 24 selected novel peptides. Validation classifications adopted in the current research (clear (panel A), unclear (panel B), and not detected (panel C)) for the 24 selected novel peptides (identified by both spectral correlations and *de novo* peptide sequencing approach) are illustrated in Figure 9. The selected transitions as well as the visual analysis for the validation of all novel peptides are shown in Figure S3. Out of 24 synthetic peptides, 16 peptides have shown very clear transition alignments. However, six peptides (peptides numbers 1, 13, 14, 15, 21 and 23) have presented less clear transition alignments, as well as a less similar pattern of peak area proportion between biological and standard injections. Yet, only two novel peptides (peptides numbers 17 and 19) could not be detected. With the retention time calibration (40 femtomolar μL^{-1} of standards spiked into a biological sample pool injection), non-existent or very small retention time shifts can be observed among the

majority of novel peptides (left and center chromatograms of each peptide's illustration) (Figure S3).

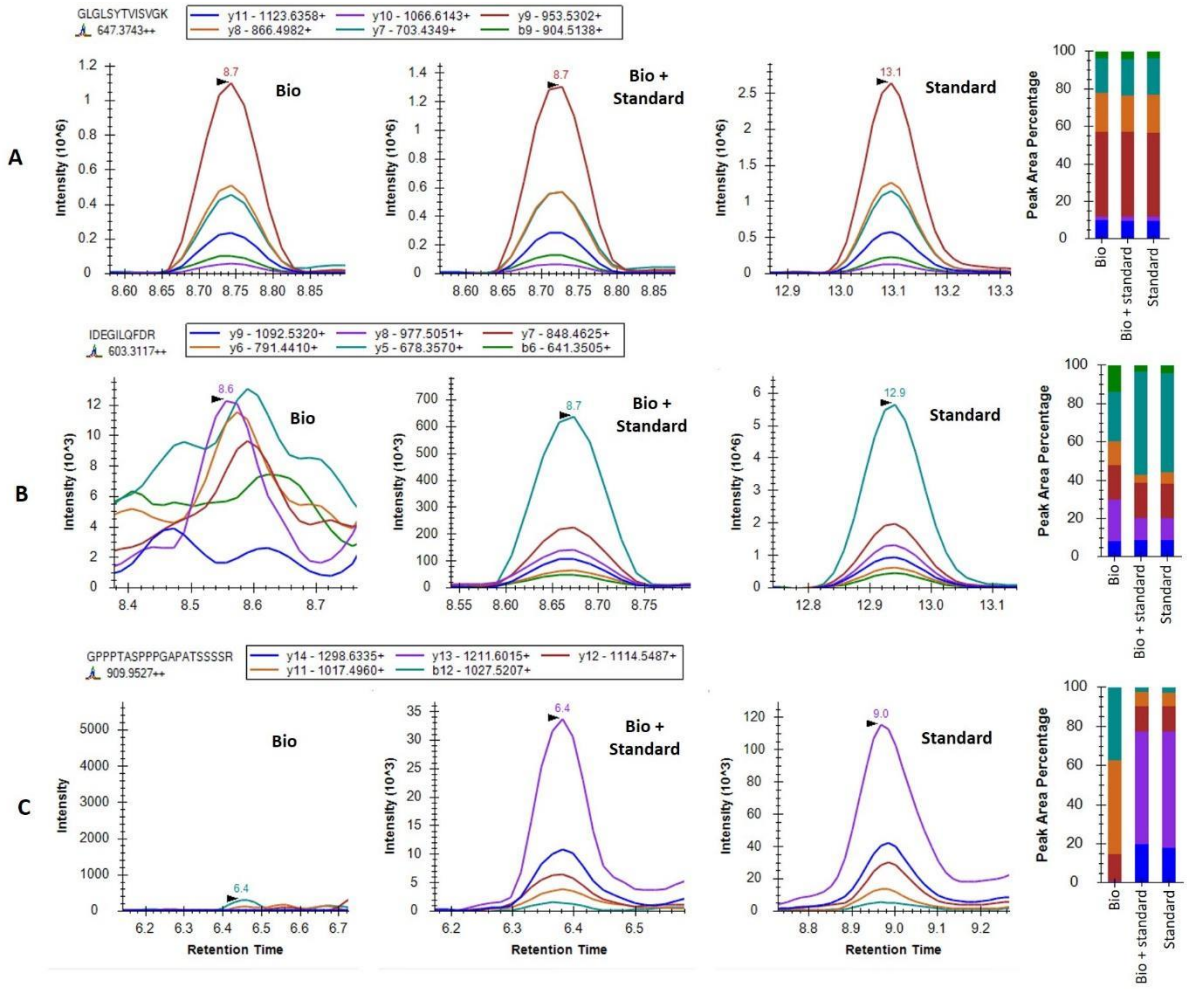


Figure 9 - Selected examples of MRM validation analysis of novel peptides identified through spectral correlations and *de novo* peptide sequences. Chromatograms are separated into: left: chromatographic traces of the monitored transitions for each peptide from a selected biological sample; center: chromatographic traces of the monitored transitions from the biological sample and a spiked-in standard concentration of 40 femtomoles μL^{-1}; right: chromatographic traces of the monitored transitions only from standard peptide at a concentration of 40 femtomolaes μL^{-1}. Bar graphs represent the raw peak area proportion between all selected transitions for each novel peptide and each color represents the signal from individual transitions. Different panels A (clear), B (unclear), and C (not detected) represent a novel peptide example of each classification.

4. DISCUSSION

4.1. Impact of seasonality and water restriction imposition on known protein abundance levels

In our study, the highest proteomic variability was identified in the summer according to the previous PCA analysis (section 3.2), which corroborates with the effect of water restriction imposition during the summer that led to a large number of proteins exhibiting significant abundance changes (mostly down-regulated). The greater proteomic variability achieved in the summer could be associated with the higher precipitation and temperature averages (especially at night) recorded throughout the summer. This variability may be attributed to *Eucalyptus* trees being more physiologically active under such conditions, which contributed to a higher proteome variability when compared to winter and spring seasons. Higher physiological performance in the summer has also been reported in other C3 tree species such as Citrus (Ribeiro *et al.*, 2009).

Lower night temperature conditions were considered the main environmental element limiting photosynthesis under a normal watering regime leading to dysfunctions in plant water status, decreased hydraulic conductance, stomatal closure, and low shoot hydration (Ribeiro *et al.*, 2009). Enrichment analysis of the most abundant known proteins over different seasons (in both watering conditions) showed that most proteins were associated with the photosynthesis process. Photosynthesis related proteins were also the main identified class among other studies involving leaf proteome and water restriction in different *Eucalyptus* species (Correia *et al.*, 2016; Martins *et al.*, 2020).

The distribution of functional classification according to COG categories of known protein groups across different seasons followed the same distribution pattern regardless of water restriction imposition. Functional classifications as post-translational modification, protein turnover, and chaperones were the main groups (highest number of known proteins identified) influenced by both seasonality (section 3.3) and water restriction imposition in the summer (section 3.4). In other studies, this functional classification group was attributed to being mainly associated with plant survival and plant abiotic stress tolerance in peanuts (Xu *et al.*, 2020), and cucumber

(Du *et al.*, 2019). Interestingly, chaperones, known to play an important role in protein folding, assembly, translocation, and membrane stabilization, can protect plants against abiotic stresses by establishing normal protein conformation and cellular homeostasis (Wang *et al.*, 2004). Yet, the unknown functional classification revealed the highest number of protein identifications throughout t-test and ANOVA-significant known protein groups, which implies the existence of a high number of proteins with yet no clear functional annotation.

4.2. Single amino acid substitution impact on protein stability

Among novel peptide identifications by both spectral correlations and *de novo* peptide sequencing approach, those identified through the latter approach pointed out a predominant response to the effect of both seasonality and water restriction imposition in *E. grandis* leaf proteome. Only two novel peptides identified by the spectral correlation approach with functional annotation described as Minichromosome maintenance and Exostosin family protein highlighted a significant effect on biological conditions in the study. Those proteins are mainly related to DNA-dependent ATPases required for initiating eukaryotic DNA replication (Coxon *et al.*, 1992) and cell wall biosynthesis (Coutinho *et al.*, 2003), respectively. In our study, single amino acid substitutions suggested a response from the plant host to reduce the protein stability under lower water availability conditions. Therefore, *Eucalyptus* trees could use SAS as a molecular response to reduce biochemical activity of photosynthesis-related proteins.

When assessing the effect of seasonality (Figure 6), the seasons with a lower rainfall pattern (winter and spring) showed a higher peptide abundance for most of the peptides with single amino acid substitutions. The highest expression rates are usually seen during the winter, which is the season of the year that has the lowest amount of rain in the experimental site located in the Northwestern region of Sao Paulo state. Its climate is classified as tropical sub-humid with dry winters, covering approximately 18.7% of the state's area (de Souza Rolim and Lucas, 2016). Only a single protein did not follow this trend, which was O-methyltransferase 1. This protein showed an increased abundance level under both the summer condition (ANOVA) and under no

water restriction imposition in the summer (t-test). In plants, the maintenance of DNA methylation relies on cytosine sequence context and is catalyzed by DNA methyltransferases, which are regulated by different mechanisms (Zhang *et al.*, 2018). DNA methylation at the 5' position of cytosine can contribute to the epigenetic regulation of nuclear gene expression and genome stability (Robertson, 2005). However, O-methyltransferases (OMTs) are a broad family of enzymes found in plants that methylate the oxygen atom in a range of secondary metabolites such as phenylpropanoids, flavonoids, and alkaloids (Lam *et al.* 2007). Assuming that O-methyltransferase 1 identified here is related to the DNA methylation process, its reduction of protein stability could lead to alterations in gene expression (Zhang *et al.*, 2018). Also, another novel peptide identified in higher abundance in the winter under water restriction imposition belonged to the ATPase, F1 complex, gamma subunit, which can be found in the mitochondria or chloroplast of plants. F1, a water-soluble portion of ATP synthase, is the site of ATP synthesis (from ADP and inorganic phosphate), while protons flow through the membrane-embedded F0 portion (Yasuda *et al.*, 2001)**.**

The majority of significant SAS identified by *de novo* peptide sequence (from both ANOVA and t-test comparisons) has consisted mainly of photosynthetic related proteins. Some enzymes were located in the stroma of chloroplasts (Ribulose 1,5-bisphosphate carboxylase/oxygenase (RuBisCO, Phosphoribulokinase, and RuBisCO activase) playing an important role in the Calvin cycle. RuBisCO and RuBisCO activase have been suggested to play an essential role in the biochemical limitations of photosynthesis rate in different C3 species (rice, wheat) as well as C4 species (maize) (Perdomo *et al.*, 2017). Assessing the proteome of Eucalyptus species (*E. saligna* and *E. tereticornis*), which have contrasting water stress tolerance, RuBisCO and RuBisCO activase (RCA) enzyme showed higher abundance in response to water stress in a tolerant species (*E. saligna*) (Martins *et al.*, 2020). Similarly, fructose-bisphosphate aldolase (FBA) and ATP synthase CF1 alpha subunit were identified in higher amounts being considered to play very important roles in increasing carbon fixation and increasing water stress tolerance, respectively (Martins *et al.*, 2020).

Important structural proteins for photosynthesis photochemical phase were identified in the Thylakoid membrane in photosystems I and II as light-harvesting

chlorophyll-proteins or photosystem reaction proteins. Likewise, fructose-bisphosphate aldolase, an essential enzyme for glycolysis and gluconeogenesis in the cytoplasm and the Calvin cycle, was identified in both ANOVA and t-test comparisons. Additionally, a SAS peptide was identified in a ubiquitous enzyme called carbonic anhydrase, which can be found in chloroplasts, mitochondria, cytosol, and plasma membrane. This enzyme plays a significant role in plants to facilitate the diffusion of CO_2 to the site of inorganic carbon fixation, and can participate in many physiological, biochemical, and structural modifications in plant metabolism (Floryszak-Wieczorek and Arasimowicz-Jelonek, 2017). Lastly, for the majority of SAS peptides, their native peptide forms were also detected under the same biological condition (mainly detected in higher abundance levels than the substituted form). We assume that this event could be associated with the presence of duplicated gene regions or genes in heterozygosity within *E. grandis* genome.

Water stress can affect cell structures and impair various biological functions, thereby resulting in the inhibition of photosynthesis, metabolic dysfunction, and damage to membranes and proteins (Krasensky and Jonak, 2012). The main effects of lack of water led to changes in the abundance of proteins involved in signal perception, defense, metabolism, and photosynthesis, suggesting a connection among those processes to avoid and/or tolerate drought (Katam *et al.*, 2016). However, plants have several mechanisms that have evolved to control water status, maintain turgor pressure, and regulate water loss (Chai *et al.*, 2016). The leaf proteome of two *E. globulus* species, under contrasting water-stressed conditions, revealed lower photosynthesis, disrupted cell growth, and alterations in gene regulation (Correia *et al.*, 2016). Another tree species, date palm (*Phoenix dactylifera*), in a study of protein abundance changes in leaves, showed protein plasticity as an important mechanism of molecular adaptation to environmental fluctuations (Ghirardo *et al.*, 2021).

As a result of single amino acid substitutions, some authors have described those events as being attributed to important functions among different organisms. A minor change in the amino acid composition of a protein may alter activity, shape, and/or function of a protein. Those alterations may lead to no impact on its functionality, improve its functionality, alter its function, or be deleterious (Abraham *et al.*, 2015). In bacteria, a single amino acid substitution in ribosomal protein (RpsU) led to a switch

between multiple-stress resistant and high fitness states (Koomen *et al.*, 2021). Besides, more than half of monogenic diseases in humans are caused by single amino acid substitutions, and a common mechanism by which amino acid substitutions cause disease is protein stability change (Teng *et al.*, 2010). In plants, a single amino acid substitution in an abscisic acid-responsive kinase (AtMPK12) drives variation in water use efficiency (WUE) between two natural genotypes of *Arabidopsis thaliana.* They resulted in physiological changes within the whole-plant transpiration efficiency reducing fitness under water-limited compared with well-watered conditions (Marais *et al.*, 2014).

The non-covalent interaction changes within protein structures were noticed by comparing native and substituted forms. Those alterations revealed an impact on the protein stability as new combinations of different interactions (hydrogen bonds, hydrophobic, ionic, and Van der Waals interactions) were created for SAS peptides, which led to a significant decrease in protein stability for nearly all cases. Such interaction changes can result in substantial protein structural modifications by altering the accumulation of hydrophobic amino acid side chains in the protein interior, changes in the protein-folding process, and limitations in the structural flexibility by conferring a uniqueness to a particular protein structure (Nelson and Cox, 2017).

In this study, no experimental analyses were carried out to assess the actual result of these single amino acid substitutions in *Eucalyptus* plant phenotype. However, the overall reduction of protein stability accomplished by estimating the quantitative change upon unfolding Gibbs free energy showed protein stability loss, which could reduce protein activity levels. Thus, those novel peptides were closely associated with tree response to water restriction adaptation by the accumulation of less stable protein forms in their leaves.

4.3. Multiple reaction monitoring validation assays

Using multiple reaction monitoring (MRM) assays with synthetic peptides provided an additional level of identification specificity for novel peptides selected for validation. Other studies have successfully shown this method as a powerful tool to

provide both identification and quantification of analytes of interest (Freue and Borchers, 2012; Percy *et al.*, 2013).

In this paper, MRM assays provided a very robust analytical performance alongside a low retention time shift between endogenous and synthetic analytes. Most of the synthetic peptides could be detected using the proposed validation method. However, some synthetic peptides presented a lower detection response since a well-defined transition alignment was only achieved under more concentrated injections (peptides 2, 6, 10, 17, 18, 19, and 22 (Figure S3)). Furthermore, novel peptides 17 and 19 were not detected in the biological samples, and this may be associated with the low abundance of such peptides within the biological sample. Thus, such concentrations were below the lower limit of detection of the mass spectrometer. Additionally, a high peak area intensity difference between the most intense and the second most intense transitions may also have imposed additional difficulties on identifying at least three aligned transitions for peptides 13, 14, and 21.

Intrinsic characteristic of each synthetic peptide may affect its specific cleavage sites within the amino acid sequence. Therefore, under a low peptide abundance scenario in the biological sample, not all of the transitions can be detected. Those lower-intensity transitions can be visually difficult to be defined when their abundance levels are similar to the background noise of the run. Likewise, due to each peptide fragmentation pattern, six transitions could not have been selected and monitored for each synthetic peptide. Also, ionic suppression (Jessome and Volver, 2006), generally related to the different ionization efficiency by competing peptides in the ionization source, could have negatively affected the identification of similar intensity transitions in the spiked-in sample for peptide 1.

5. CONCLUSIONS

The highest protein variability was achieved in the summer, and most known proteins were down-regulated under such conditions when water restriction was present. Photosynthesis-related proteins were predominant across seasons regardless of water restriction imposition with the highest abundance levels. Furthermore, water restriction imposition and drier seasons (Winter and Spring) led to a higher expression of less stable SAS peptides mainly related to the photosynthesis process. Those single amino acid substitutions could work as a plant physiological mechanism to cope with abiotic stresses (water). Further experimental studies on protein function and activity may clarify the role of each amino acid substitution within the plant to withstand different environmental conditions and water restrictions.

6. REFERENCES

Abraham PE, Wang X, Ranjan P, Nookaew I, Zhang B, Tuskan GA, Hettich RL (2015) Integrating mRNA and protein sequencing enables the detection and quantitative profiling of natural protein sequence variants of *Populus trichocarpa*. **Journal of Proteome Research** 14: 5318–5326.

Alvares CA, Stape JL, Sentelhas PC, de Moraes Gonçalves JL, Sparovek G (2013) Köppen's climate classification map for Brazil. **Meteorologische Zeitschrift** 22: 711–728.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. **Nature Genetics** 25(1): 25-9.

Bereman MS, Maclean B, Tomazela DM, Liebler DC, Maccoss MJ (2012) The development of selected reaction monitoring methods for targeted proteomics via empirical refinement. **Proteomics** 12: 1134–1141.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The Protein Data Bank. **Nucleic Acids Research** 28(1): 235-42.

Binkley D, Campoe OC, Alvares C, Carneiro RL, Cegatta Í, Stape JL (2017) The interactions of climate, spacing and genetics on clonal *Eucalyptus* plantations across Brazil and Uruguay. **Forest Ecology and Management** 405: 271–283.

Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J (2021) eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. **Molecular Biology and Evolution** 9;38(12): 5825-5829.

Chai Q, Gan Y, Zhao C, Xu HL, Waskom RM, Niu Y, Siddique KHM (2016) Regulated deficit irrigation for crop production under drought stress. A review. **Agronomy for Sustainable Development** 36: 1–21.

Chen Y, Lu H, Zhang N, Zhu Z, Wang S, Li M (2020) PremPS: Predicting the impact of missense mutations on protein stability. **PLoS Computational Biology** 16(12): e1008543. doi: 10.1371/journal.pcbi.1008543

Cheng J, Randall A, Baldi P (2006) Prediction of protein stability changes for single-site mutations using support vector machines. **Proteins: Structure, Function and Genetics** 62: 1125–1132.

Correia B, Valledor L, Hancock RD, Renaut J, Pascual J, Soares AMVM, Pinto G (2016) Integrated proteomics and metabolomics to unlock global and clonal responses of *Eucalyptus globulus* recovery from water deficit. **Metabolomics** 12: 141. doi: 10.1007/s11306-016-1088-4

Coutinho PM, Stam M, Blanc E, Henrissat B (2003) Why are there so many carbohydrate-active enzyme-related genes in plants? **Trends in Plant Science** 8: 563–565.

Cox J, Hein MY, Luber CA, Paron I, Nagaraj N, Mann M (2014) Accurate Proteome-wide Label-free Quantification by Delayed Normalization and Maximal Peptide Ratio Extraction, Termed MaxLFQ* □ S Technological Innovation and Resources. **Molecular & Cellular Proteomics** 13: 2513–2526.

Cox J, Mann M (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. **Nature Biotechnology** 26: 1367–1372.

Cox J, Neuhauser N, Michalski A, Scheltema RA, Olsen J v., Mann M (2011) Andromeda: A peptide search engine integrated into the MaxQuant environment. **Journal of Proteome Research** 10: 1794–1805.

Coxon A, Maundrell' K, Kearsey SE (1992) Fission yeast cdc2l+ belongs to a family of proteins involved in an early step of chromosome replication. **Nucleic Acids Research** 20(21): 5571–5577.

Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N, Rokhsar DS (2012) Phytozome: a comparative platform for green plant genomics, **Nucleic Acids Research** 40 (D1): D1178-D1186.

Du C, Chai L, Wang Z, Fan H (2019) Response of proteome and morphological structure to short-term drought and subsequent recovery in *Cucumis sativus* leaves. **Physiologia Plantarum** 167: 676–689.

Easterling DR, Meehl GA, Parmesan C, Changnon SA, Karl TR, Mearns LO (2000) Climate Extremes: Observations, Modeling, and Impacts. **Science** 289 (5487): 2068-2074.

Eng JK, Mccormack AL, Yates JR (1994) An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database. **Journal of the American Society for Mass Spectrometry** 11: 976-89.

Feist P, Hummon AB (2015) Proteomic challenges: Sample preparation techniques for Microgram-Quantity protein analysis from biological samples. **International Journal of Molecular Sciences** 16: 3537–3563.

Floryszak-Wieczorek J, Arasimowicz-Jelonek M (2017) The multifunctional face of plant carbonic anhydrase. **Plant Physiology and Biochemistry** 112: 362–368.

Freeman JS, Potts BM, Downes GM, Pilbeam D, Thavamanikumar S, Vaillancourt RE (2013) Stability of quantitative trait loci for growth and wood properties across multiple pedigrees and environments in *Eucalyptus globulus*. **New Phytologist** 198: 1121–1134.

Freue GVC, Borchers CH (2012) Multiple Reaction Monitoring (MRM): Principles and application to coronary artery disease. **Circulation: Cardiovascular Genetics** 3: 378. doi: 10.1161/CIRCGENETICS.111.959528

Ghirardo A, Nosenko T, Kreuzwieser J, Winkler JB, Kruse J, Albert A, Merl-Pham J, Lux T, Ache P, Zimmer I, et al (2021) Protein expression plasticity contributes to heat and drought tolerance of date palm. **Oecologia** 197: 903–919.

Han Y, Ma B, Zhang K (2005) SPIDER: Software for protein identification from sequence tags with *de novo* sequencing Error. **Journal of Bioinformatics and Computational Biology** 3 :697-716.

Houston NL, Lee DG, Stevenson SE, Ladies GS, Bannon GA, McClain S, Privalle L, Stagg N, Herouet-Guieheney C, MacIntosh SC, et al (2011) Quantitation of soybean allergens using tandem mass spectrometry. **Journal of Proteome Research** 10: 763–773.

IBÁ - Indústria Brasileira de Árvores (2020) **Annual Report**. Brasília, 66 p.

Jaffe JD, Berg HC, Church GM (2004) Proteogenomic mapping as a complementary method to perform genome annotation. **Proteomics** 4: 59–77.

Jessome LL, Volmer DA (2006) Ion suppression: a major concern in mass spectrometry. **LCGC North America** 24(5): 498−510.

Jorge GL, Balbuena TS (2021) Identification of novel protein-coding sequences in *Eucalyptus grandis* plants by high-resolution mass spectrometry. **Biochimica et Biophysica Acta - Proteins and Proteomics** 1869(3): 140594. doi: 10.1016/j.bbapap.2020.140594

Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. **Nucleic Acids Research** 28(1): 27-30.

Katam R, Sakata K, Suravajhala P, Pechan T, Kambiranda DM, Naik KS, Guo B, Basha SM (2016) Comparative leaf proteomics of drought-tolerant and -susceptible peanut in response to water stress. **Journal of Proteomics** 143: 209–226.

Kim D, Yang J, Gu F, Park S, Combs J, Adams A, Mayes HB, Jeon SJ, Bahk JD, Nielsen E (2021) A temperature-sensitive FERONIA mutant allele that alters root hair growth. **Plant Physiology** 185: 405–423.

Koomen J, Huijboom L, Ma X, Tempelaars MH, Boeren S, Zwietering MH, den Besten HMW, Abee T (2021) Amino acid substitutions in ribosomal protein RpsU enable switching between high fitness and multiple-stress resistance in *Listeria monocytogenes.* **International Journal of Food Microbiology** 351: 109269. doi: 10.1016 /j.ijfoodmicro.2021.109269

Krasensky J, Jonak C (2012) Drought, salt, and temperature stress-induced metabolic rearrangements and regulatory networks. **Journal of Experimental Botany** 63: 1593–1608.

Kucukkal TG, Petukh M, Li L, Alexov E (2015) Structural and physico-chemical effects of disease and non-disease nsSNPs on proteins. **Current Opinion in Structural Biology** 32: 18–24.

Lam KC, Ibrahim RK, Behdad B, Dayanandan S (2007) Structure, function, and evolution of plant O-methyltransferases. **Genome** 50(11): 1001-1013.

Li H, Li Y, Ke Q, Kwak SS, Zhang S, Deng X (2020) Physiological and differential proteomic analyses of imitation drought stress response in *Sorghum bicolor* root at the seedling stage. **International Journal of Molecular Sciences** 21: 1–28.

Li J, Su Z, Ma Z-Q, C Slebos RJ, Halvey P, Tabb DL, Liebler DC, Pao W, Zhang B (2011) A bioinformatics workflow for variant peptide detection in shotgun Proteomics. **Molecular & Cellular Proteomics** 10(5): M110.006536.

MacLean B, Tomazela DM, Shulman N, Chambers M, Finney GL, Frewen B, Kern R, Tabb DL, Liebler DC, MacCoss MJ (2010) Skyline: An open-source document editor for creating and analyzing targeted proteomics experiments. **Bioinformatics** 26: 966–968.

Marais DLD, Auchincloss LC, Sukamtoh E, McKay JK, Logan T, Richards JH, Juenger TE (2014) Variation in MPK12 affects water use efficiency in *Arabidopsis* and reveals a pleiotropic link between guard cell size and ABA response. **Proceedings of the National Academy of Sciences** 111: 2836–2841.

Martínez-Márquez A, Morante-Carriel J, Sellés-Marchart S, Martínez-Esteso MJ, Pineda-Lucas JL, Luque I, Bru-Martínez R (2013) Development and validation of MRM methods to quantify protein isoforms of polyphenol oxidase in loquat fruits. **Journal of Proteome Research** 12: 5709–5722.

Martins R de S, Faria JMR, Rossini BC, Marino CL, dos Santos LD, José AC (2020a) Proteomic analyses unraveling water stress response in two *Eucalyptus* species originating from contrasting environments for aridity. **Molecular Biology Reports** 47: 5191–5205.

Nelson DL, Cox MM (2017) **Princípios de Bioquímica de Lehninger**, 7th ed. W.H. Freeman, 1312 p.

Nesvizhskii AI (2014) Proteogenomics: Concepts, applications and computational strategies. **Nature Methods** 11: 1114–1125.

Percy AJ, Chambers AG, Yang J, Borchers CH (2013) Multiplexed MRM-based quantitation of candidate cancer biomarker proteins in undepleted and non-enriched human plasma. **Proteomics** 13: 2202–2215.

Perdomo JA, Capó-Bauçà S, Carmo-Silva E, Galmés J (2017) Rubisco and rubisco activase play an important role in the biochemical limitations of photosynthesis in rice, wheat, and maize under high temperature and water deficit. **Frontiers in Plant Science** 8: 490. doi: 10.3389/fpls.2017.00490

Perez-Riverol Y, Bai J, Bandla C, García-Seisdedos D, Hewapathirana S, Kamatchinathan S, Kundu DJ, Prakash A, Frericks-Zipper A, Eisenacher M, et al (2022) The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. **Nucleic Acids Research** 50: D543–D552.

Pires DEV, Ascher DB, Blundell TL (2014) MCSM: Predicting the effects of mutations in proteins using graph-based signatures. **Bioinformatics** 30: 335–342.

Ribeiro R, Machado E, Santos M, Oliveira R (2009) Photosynthesis and water relations of well-watered orange plants as affected by winter and summer conditions. **Photosynthetica** 47: 215–222.

Robertson KD (2005) DNA methylation and human disease. **Nature Reviews Genetics** 6: 597–610.

Shao HB, Chu LY, Jaleel CA, Manivannan P, Panneerselvam R, Shao MA (2009) Understanding water deficit stress-induced changes in the basic metabolism of higher plants-biotechnologically and sustainably improving agriculture and the ecoenvironment in arid regions of the globe. **Critical Reviews in Biotechnology** 29: 131–151.

Shi T, Song E, Nie S, Rodland KD, Liu T, Qian WJ, Smith RD (2016) Advances in targeted proteomics and applications to biomedical research. **Proteomics** 16: 2160–2182.

de Souza Rolim G, Lucas LE (2016) Camargo, Köppen and Thornthwaite climate classification systems in defining climatical regions of the state of São Paulo, Brazil. **International Journal of Climatology** 36: 636–643.

Steinegger M, Söding J (2017) MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. **Nature Biotechnology** 35(11): 1026-1028.

Stevenson SE, Woods CA, Hong B, Kong X, Thelen JJ, Ladics GS (2012) Environmental effects on allergen levels in commercially grown non-genetically modified soybeans: Assessing variation across North America. **Frontiers in Plant Science** 3: 196. doi: 10.3389/fpls.2012.00196

Teng S, Srivastava AK, Schwartz CE, Alexov E, Wang L (2010) Structural assessment of the effects of Amino Acid Substitutions on protein stability and protein-protein interaction. **International Journal of Computational Biology and Drug Design** 3: 334–349.

Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA (2003) The COG database: an updated version includes eukaryotes. **BMC Bioinformatics** 11;4: 41.

Tyanova S, Temu T, Cox J (2016a) The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. **Nature Protocols** 11: 2301–2319.

Tyanova S, Temu T, Sinitcyn P, Carlson A, Hein MY, Geiger T, Mann M, Cox J (2016b) The Perseus computational platform for comprehensive analysis of (prote)omics data. **Nature Methods** 13: 731–740.

Wang W, Vinocur B, Shoseyov O, Altman A (2004) Role of plant heat-shock proteins and molecular chaperones in the abiotic stress response. **Trends in Plant Science** 9: 244–252.

Xu Y, Zhang GC, Ding H, Ci DW, Qin FF, Zhang ZM, Dai LX (2020) Effects of salt and drought stresses on rhizosphere soil bacterial community structure and peanut yield. **Ying Yong Sheng Tai Xue Bao** 4: 1305-1313. doi: 10.13287/j.1001-9332.202004.036.

Yasuda R, Noji H, Yoshida M, Kinosita Jr K, Itoh H (2001) Resolution of distinct rotational substeps by submillisecond kinetic analysis of F 1-ATPase. **Nature** 410: 898–904.

Zhang H, Lang Z, Zhu JK (2018) Dynamics and function of DNA methylation in plants. **Nature Reviews Molecular Cell Biology** 19: 489–506.

Zhang J, Xin L, Shan B, Chen W, Xie M, Yuen D, Zhang W, Zhang Z, Lajoie GA, Ma B (2012) PEAKS DB: De novo sequencing assisted database search for sensitive and accurate peptide identification. **Molecular and Cellular Proteomics** 4: M111.010587. doi: 10.1074/mcp.M111.010587

APPENDIX - Supplementary Images

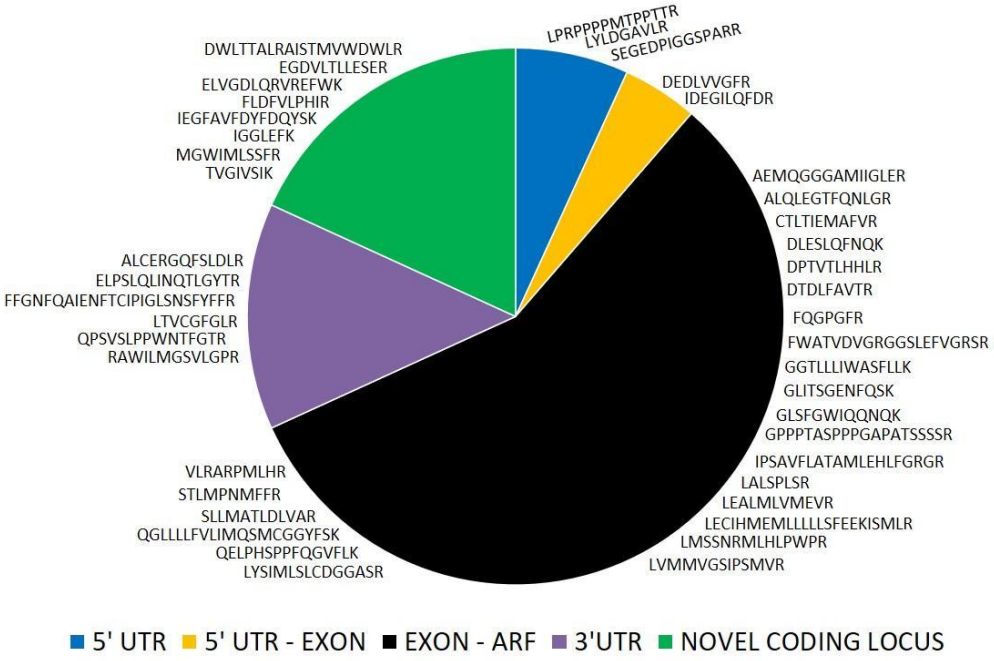


Figure S1 – Genomic classification of novel peptides identified by spectral correlations against the modified RNA databases (mapping details can be found in Supplementary Table 4). Mapping classification of the novel peptides was categorized as peptide mapped to a 5′ untranslated region (5’ UTR); peptide mapped to regions between the coding sequence and the neighboring 5’ UTR region (Exon boundary) (5’ UTR–EXON); peptide mapped to an alternative reading frame sequence within an exon region (EXON-ARF); peptide mapped to a 3′ untranslated region (3’ UTR), and peptide mapped to a novel coding locus (NCL).

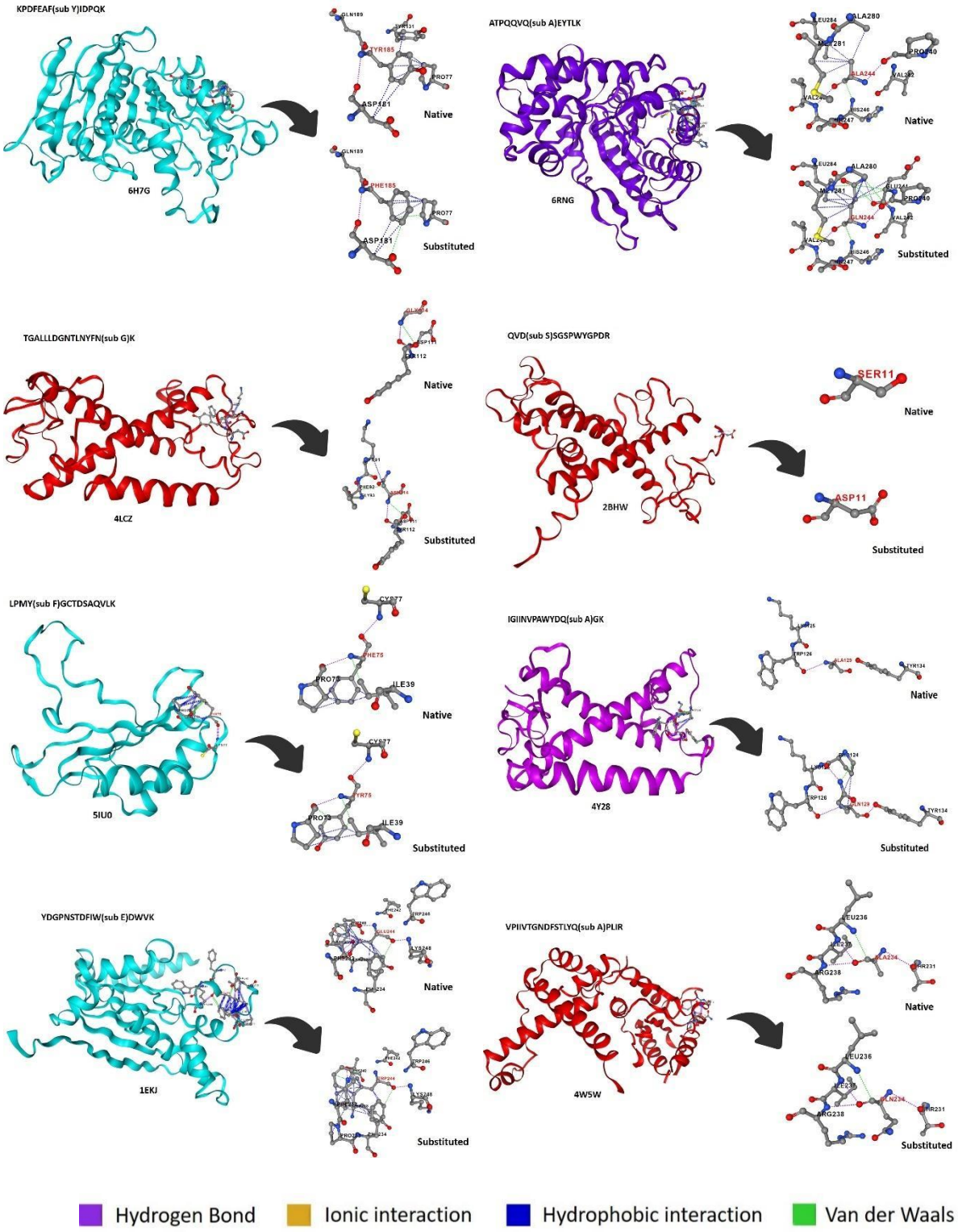
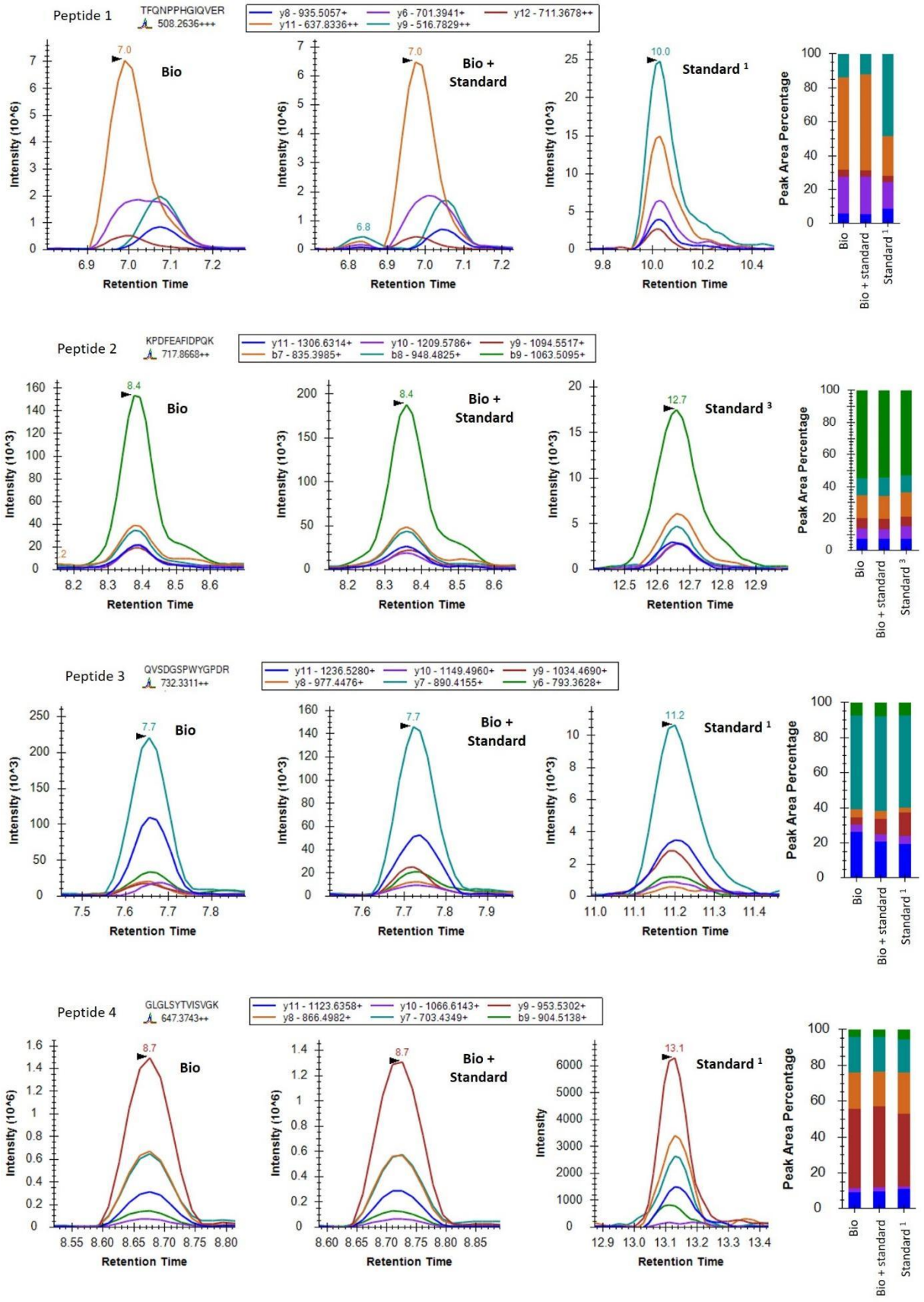
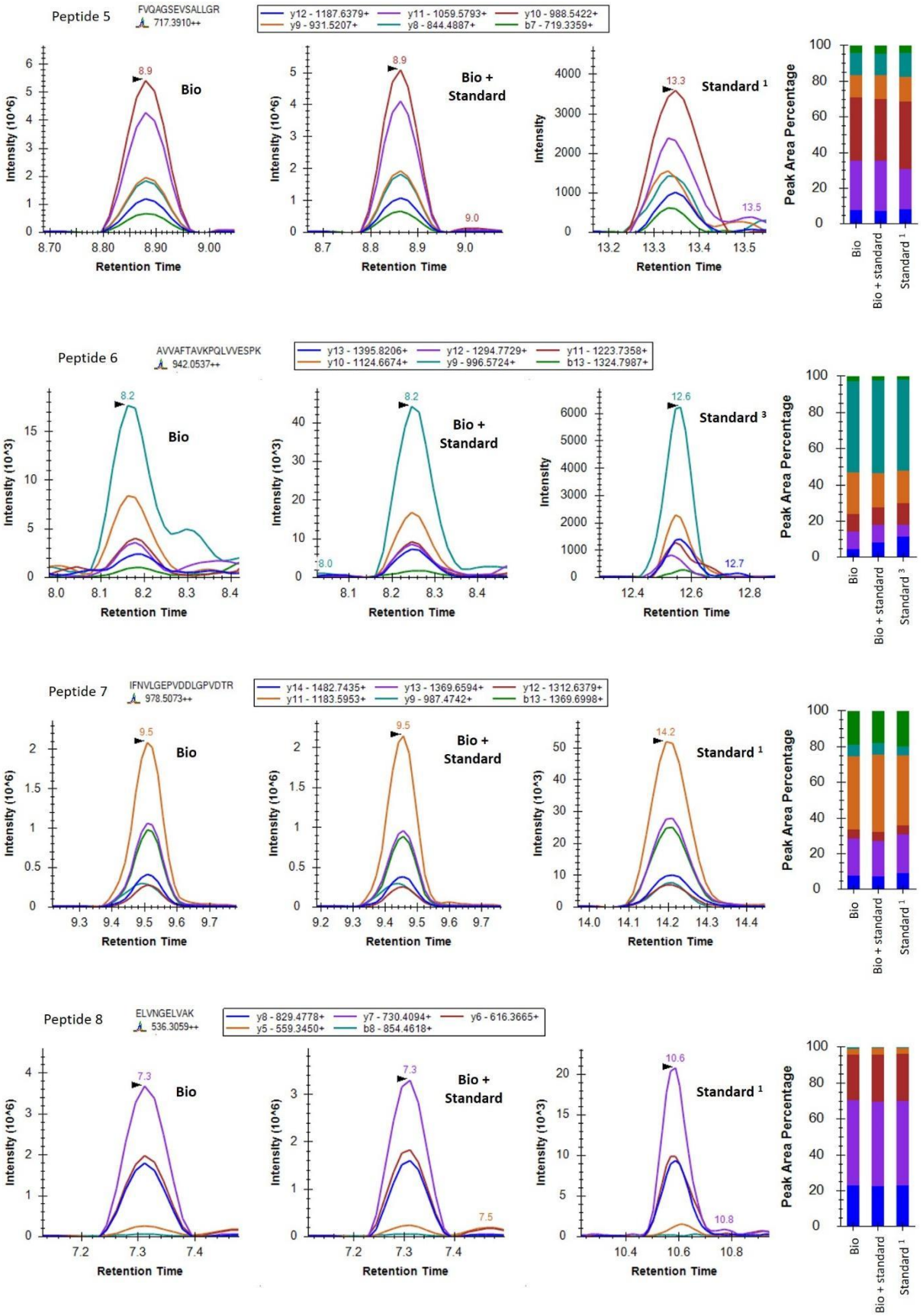
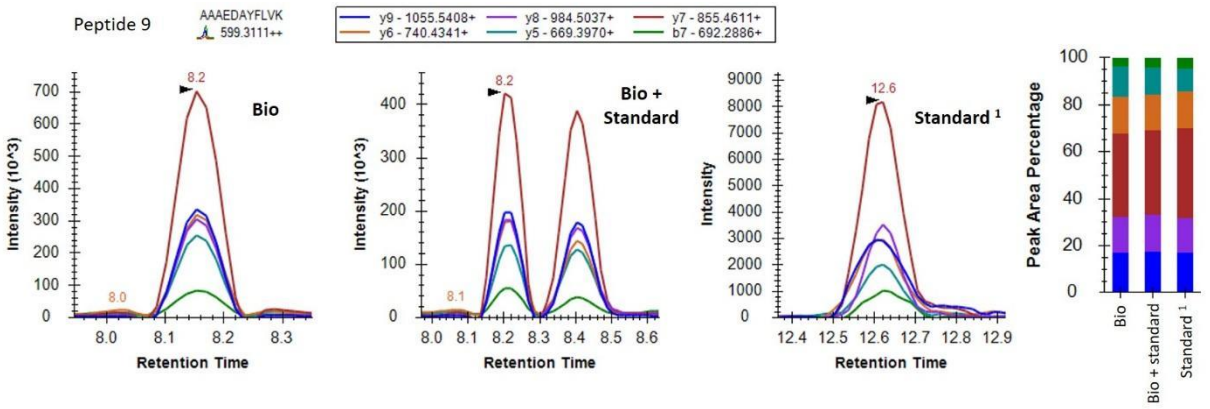
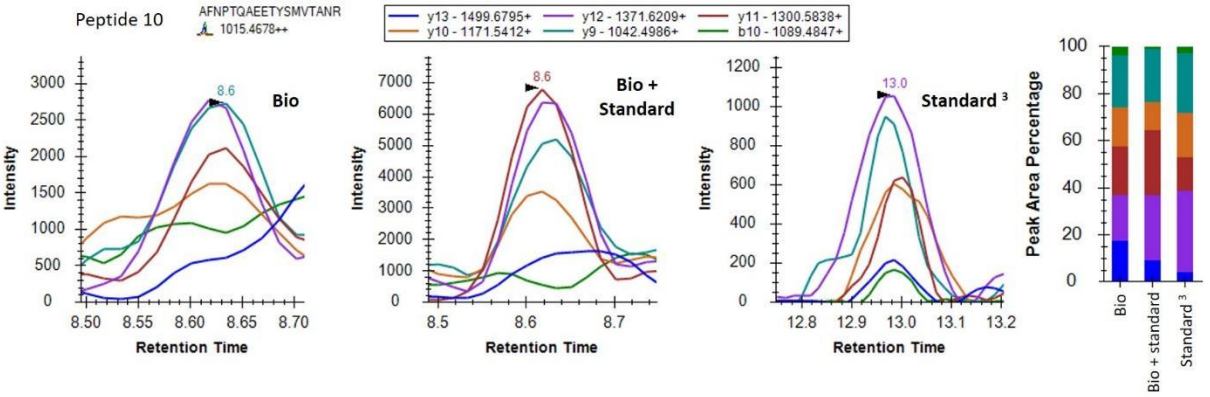


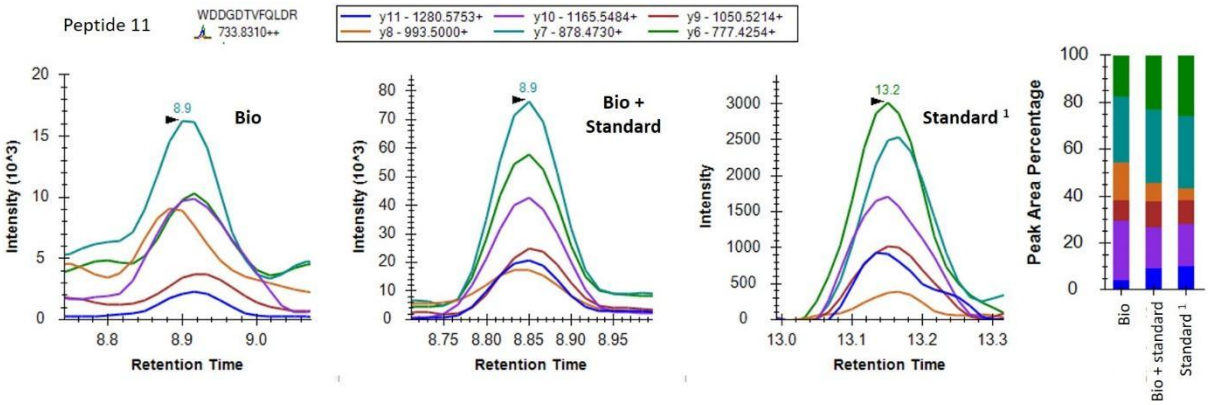
Figure S2 - Comparison of non-covalent interactions formed by a single amino acid substitution residue for the native and substituted protein forms. Novel peptide sequences are shown above each reference PDB protein structure.

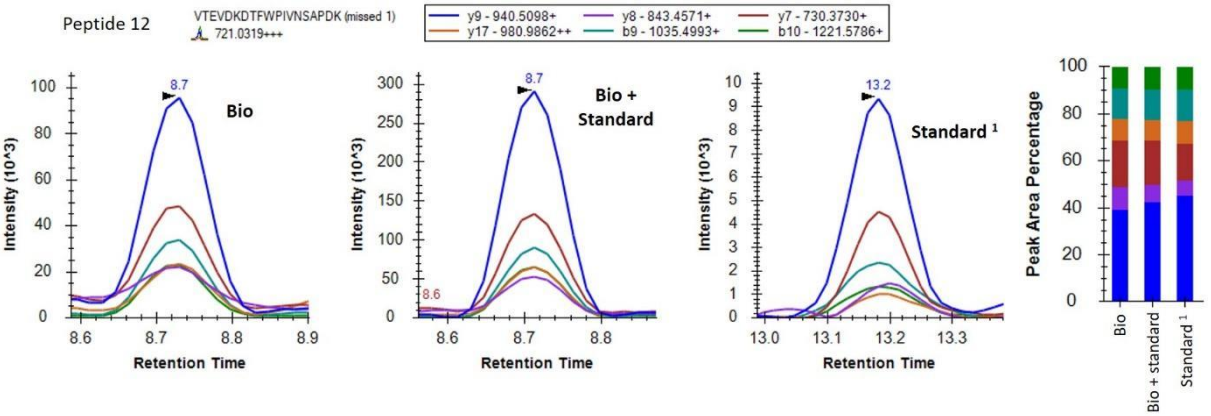

Peptide 1
TFQNPPHGIQVER
508.2636+++
y8 - 935.5057+
y6 - 701.3941+
y12 - 711.3678++
y11 - 637.8336++
y9 - 516.7829++
Bio
Bio + Standard
Standard 1
Intensity (10^6)
Intensity (10^3)
Retention Time
Peak Area Percentage
Bio
Bio + standard
Standard 1
Peptide 2
KPDFEAFIDPQK
717.8668++
y11 - 1306.6314+
y10 - 1209.5786+
y9 - 1094.5517+
b7 - 835.3985+
b8 - 948.4825+
b9 - 1063.5095+
Bio
Bio + Standard
Standard 3
Bio
Bio + standard
Standard 3
Peptide 3
QVSDGSPWYGPDR
732.3311++
y11 - 1236.5280+
y10 - 1149.4960+
y9 - 1034.4690+
y8 - 977.4476+
y7 - 890.4155+
y6 - 793.3628+
Bio
Bio + Standard
Standard 1
Bio
Bio + standard
Standard 1
Peptide 4
GLGLSYTVISVGK
647.3743++
y11 - 1123.6358+
y10 - 1066.6143+
y9 - 953.5302+
y8 - 866.4982+
y7 - 703.4349+
b9 - 904.5138+
Bio
Bio + Standard
Standard 1
Bio
Bio + standard
Standard 1

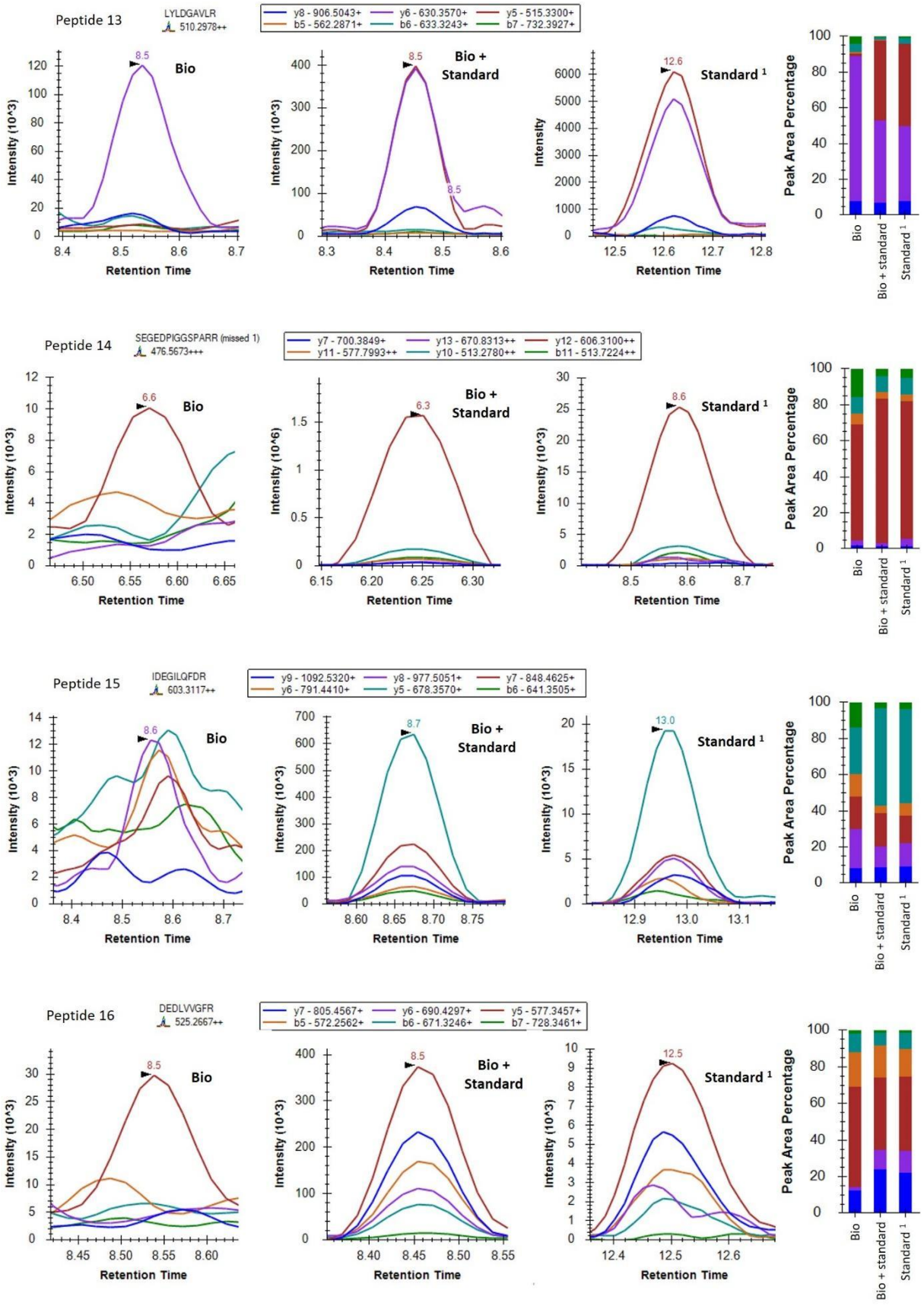

Peptide 5
FVQAGSEVSALLGR
717.3910++
y12 - 1187.6379+
y11 - 1059.5793+
y10 - 988.5422+
y9 - 931.5207+
y8 - 844.4887+
b7 - 719.3359+
Bio
Bio + Standard
Standard 1
Intensity (10^6)
Intensity
Retention Time
Peak Area Percentage
Bio
Bio + standard
Standard 1
Peptide 6
AVVAFTAVKPQLVVESPK
942.0537++
y13 - 1395.8206+
y12 - 1294.7729+
y11 - 1223.7358+
y10 - 1124.6674+
y9 - 996.5724+
b13 - 1324.7987+
Bio
Bio + Standard
Standard 3
Intensity (10^3)
Retention Time
Peak Area Percentage
Bio
Bio + standard
Standard 3
Peptide 7
IFNVLGEPVDDLGPVDTR
978.5073++
y14 - 1482.7435+
y13 - 1369.6594+
y12 - 1312.6379+
y11 - 1183.5953+
y9 - 987.4742+
b13 - 1369.6998+
Bio
Bio + Standard
Standard 1
Intensity (10^6)
Intensity (10^3)
Retention Time
Peak Area Percentage
Bio
Bio + standard
Standard 1
Peptide 8
ELVNGELVAK
536.3059++
y8 - 829.4778+
y7 - 730.4094+
y6 - 616.3665+
y5 - 559.3450+
b8 - 854.4618+
Bio
Bio + Standard
Standard 1
Intensity (10^6)
Intensity (10^3)
Retention Time
Peak Area Percentage
Bio
Bio + standard
Standard 1

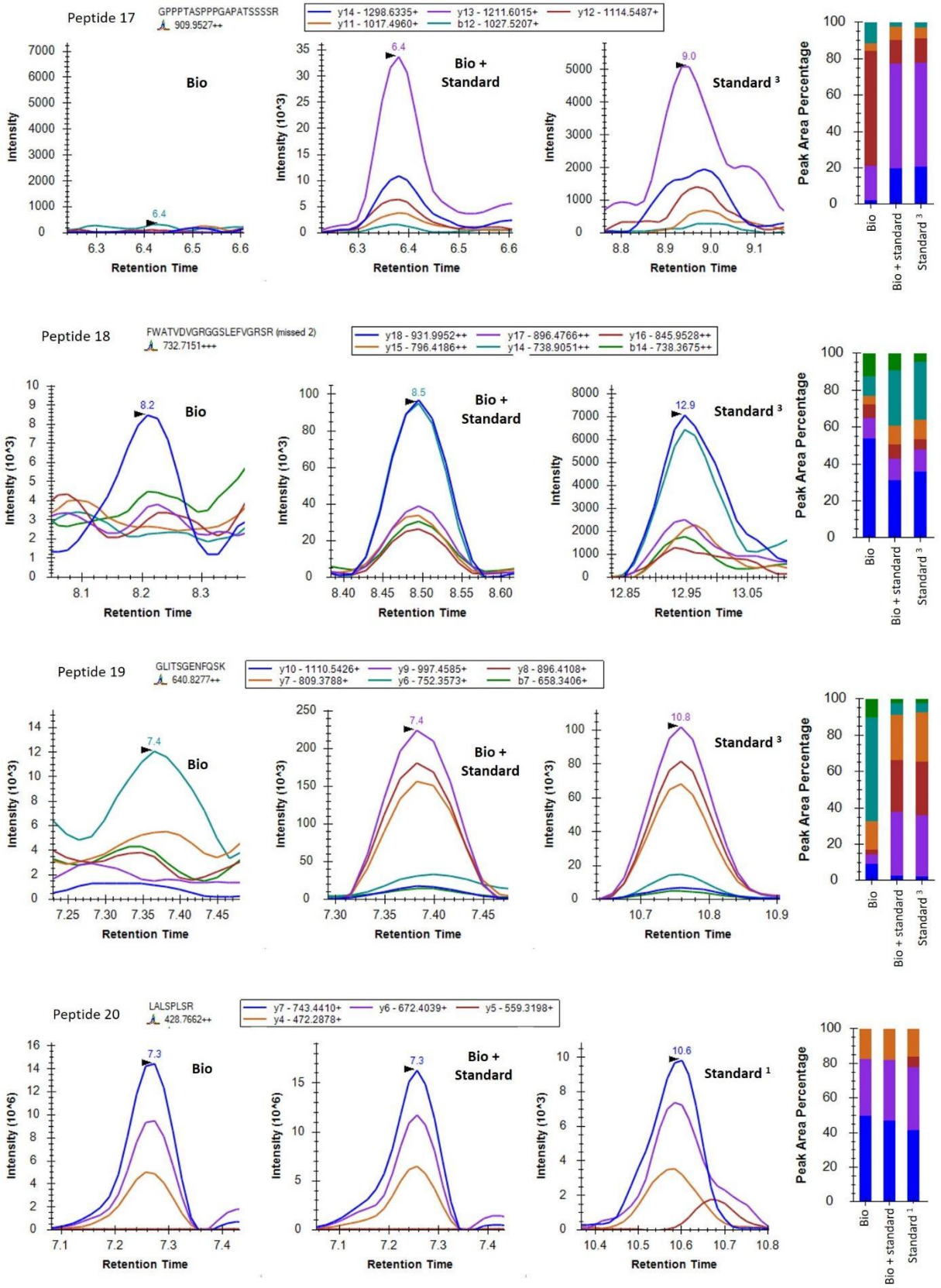

Peptide 9
AAAEDAYFLVK
599.3111++
y9 - 1055.5408+
y8 - 984.5037+
y7 - 855.4611+
y6 - 740.4341+
y5 - 669.3970+
b7 - 692.2886+
Bio
Bio + Standard
Standard 1
Intensity (10^3)
Intensity
Retention Time
Peak Area Percentage
Bio
Bio + standard
Standard 1

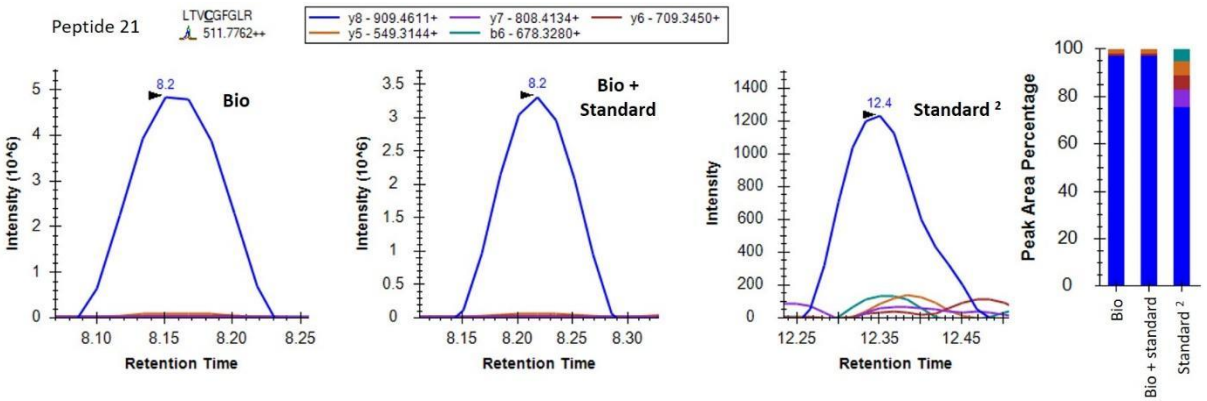

Peptide 10
AFNPTQAEETYSMVTANR
1015.4678++
y13 - 1499.6795+
y12 - 1371.6209+
y11 - 1300.5838+
y10 - 1171.5412+
y9 - 1042.4986+
b10 - 1089.4847+
Bio
Bio + Standard
Standard 3
Intensity
Retention Time
Peak Area Percentage
Bio
Bio + standard
Standard 3

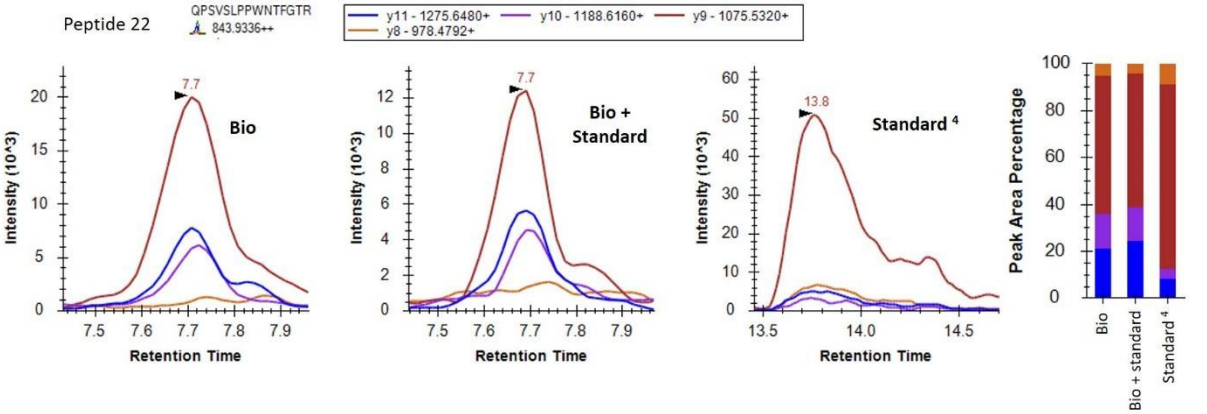

Peptide 11
WDDGDTVFQLDR
733.8310++
y11 - 1280.5753+
y10 - 1165.5484+
y9 - 1050.5214+
y8 - 993.5000+
y7 - 878.4730+
y6 - 777.4254+
Bio
Bio + Standard
Standard 1
Intensity (10^3)
Intensity
Retention Time
Peak Area Percentage
Bio
Bio + standard
Standard 1

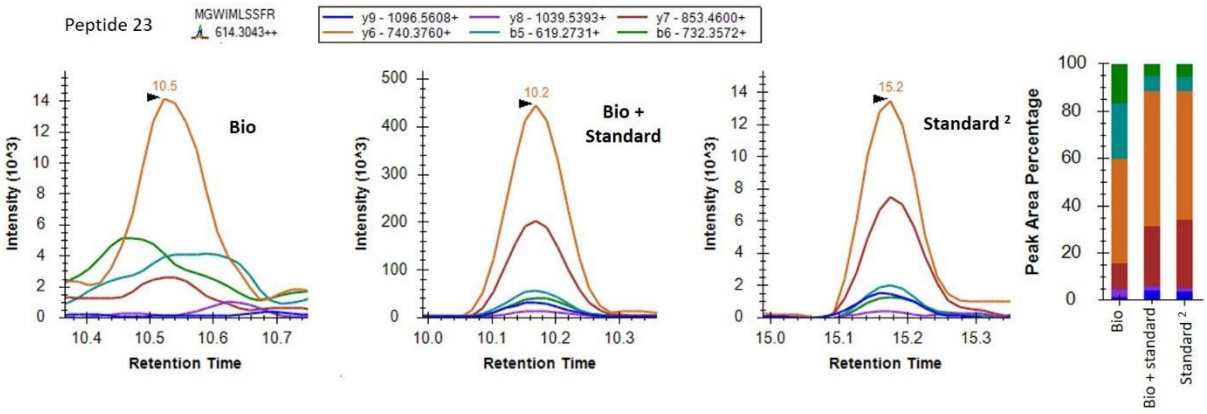

Peptide 12
VTEVDKDTFWPIVNSAPDK (missed 1)
721.0319+++
y9 - 940.5098+
y8 - 843.4571+
y7 - 730.3730+
y17 - 980.9862++
b9 - 1035.4993+
b10 - 1221.5786+
Bio
Bio + Standard
Standard 1
Intensity (10^3)
Intensity
Retention Time
Peak Area Percentage
Bio
Bio + standard
Standard 1


Peptide 13
LYLDGAVLR
510.2978++
y8 - 906.5043+
y6 - 630.3570+
y5 - 515.3300+
b5 - 562.2871+
b6 - 633.3243+
b7 - 732.3927+
Bio
Bio + Standard
Standard 1
Intensity (10^3)
Intensity
Retention Time
Peak Area Percentage
Bio
Bio + standard
Standard 1
Peptide 14
SEGEDPIGGSPARR (missed 1)
476.5673+++
y7 - 700.3849+
y13 - 670.8313++
y12 - 606.3100++
y11 - 577.7993++
y10 - 513.2780++
b11 - 513.7224++
Intensity (10^6)
Peptide 15
IDEGILQFDR
603.3117++
y9 - 1092.5320+
y8 - 977.5051+
y7 - 848.4625+
y6 - 791.4410+
y5 - 678.3570+
b6 - 641.3505+
Peptide 16
DEDLVVGFR
525.2667++
y7 - 805.4567+
y6 - 690.4297+
y5 - 577.3457+
b5 - 572.2562+
b6 - 671.3246+
b7 - 728.3461+


Peptide 17
GPPPTASPPPGAPATSSSSR
909.9527++
y14 - 1298.6335+
y13 - 1211.6015+
y12 - 1114.5487+
y11 - 1017.4960+
b12 - 1027.5207+
Bio
Bio + Standard
Standard 3
Intensity
Intensity (10^3)
Retention Time
Peak Area Percentage
Bio
Bio + standard
Standard 3
Peptide 18
FWATVDVGRGGSLEFVGRSR (missed 2)
732.7151+++
y18 - 931.9952++
y17 - 896.4766++
y16 - 845.9528++
y15 - 796.4186++
y14 - 738.9051++
b14 - 738.3675++
Peptide 19
GLITSGENFQSK
640.8277++
y10 - 1110.5426+
y9 - 997.4585+
y8 - 896.4108+
y7 - 809.3788+
y6 - 752.3573+
b7 - 658.3406+
Peptide 20
LALSPLSR
428.7662++
y7 - 743.4410+
y6 - 672.4039+
y5 - 559.3198+
y4 - 472.2878+
Standard 1
Intensity (10^6)


Peptide 21
LTVCGFGLR
511.7762++
y8 - 909.4611+
y7 - 808.4134+
y6 - 709.3450+
y5 - 549.3144+
b6 - 678.3280+
Bio
Bio + Standard
Standard 2
Intensity (10^6)
Intensity
Retention Time
Peak Area Percentage
Bio
Bio + standard
Standard 2


Peptide 22
QPSVSLPPWNTFGTR
843.9336++
y11 - 1275.6480+
y10 - 1188.6160+
y9 - 1075.5320+
y8 - 978.4792+
Bio
Bio + Standard
Standard 4
Intensity (10^3)
Retention Time
Peak Area Percentage
Bio
Bio + standard
Standard 4


Peptide 23
MGWIMLSSFR
614.3043++
y9 - 1096.5608+
y8 - 1039.5393+
y7 - 853.4600+
y6 - 740.3760+
b5 - 619.2731+
b6 - 732.3572+
Bio
Bio + Standard
Standard 2
Intensity (10^3)
Retention Time
Peak Area Percentage
Bio
Bio + standard
Standard 2

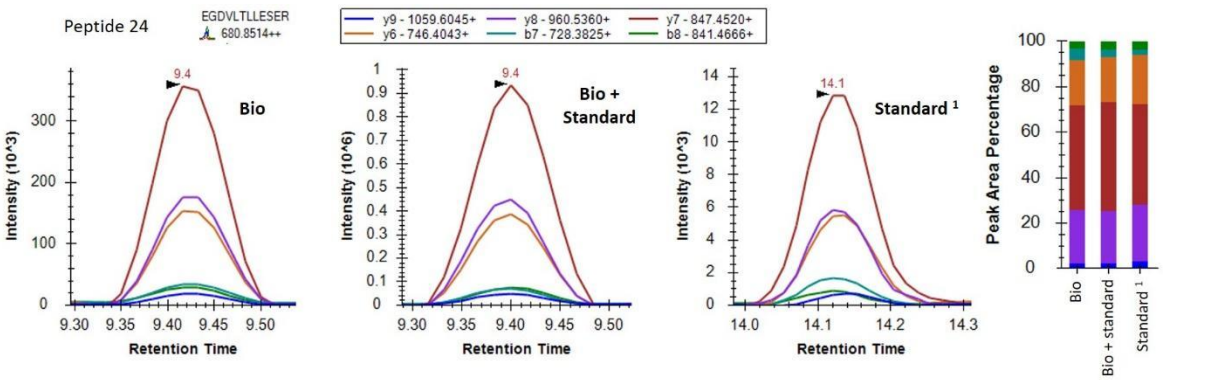
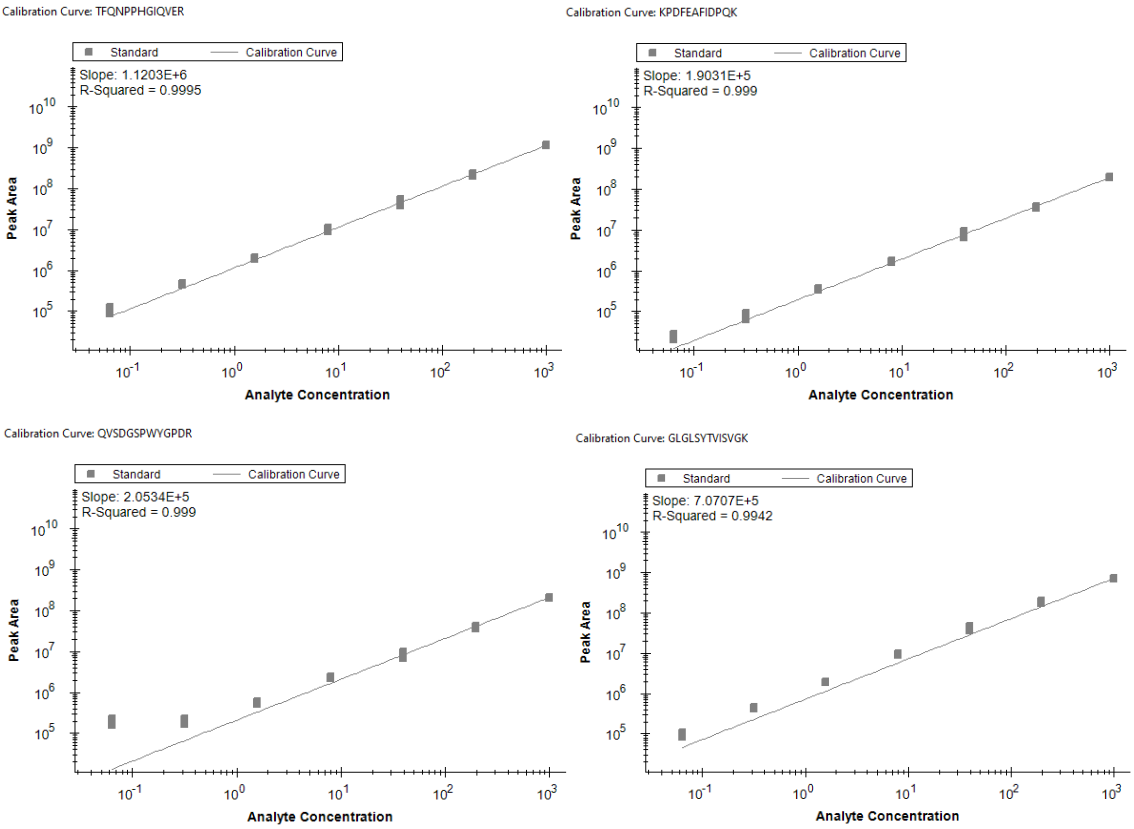
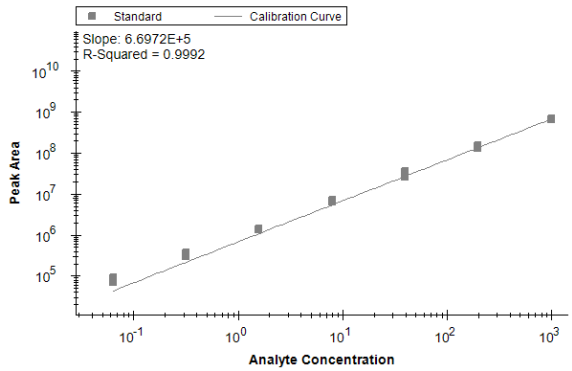

Peptide 24
EGDVLTLLESER
680.8514++
y9 - 1059.6045+
y8 - 960.5360+
y7 - 847.4520+
y6 - 746.4043+
b7 - 728.3825+
b8 - 841.4666+
Bio
Bio + Standard
Standard 1
Intensity (10^3)
Intensity (10^6)
Retention Time
Peak Area Percentage
Bio
Bio + standard
Standard 1

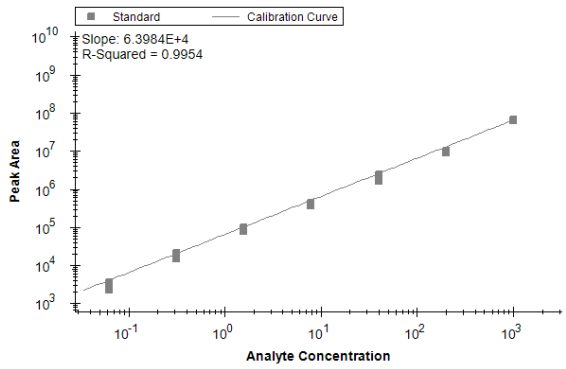
Figure S3- MRM validation analysis of 24 novel peptides identified through spectral correlations and *de novo* peptide sequences. Chromatograms are divided into: left: chromatographic traces of the monitored transitions for each peptide from a selected biological sample; center: chromatographic traces of the monitored transitions from the biological sample and a spiked-in standard concentration of 40 femtomolar μL^{-1}; right: chromatographic traces of the monitored transitions from the standard peptide at a concentration of 40 femtomolar μL^{-1}. Bar graphs represent the peak area proportion between all selected transitions for each novel peptide. [1,2,3,4] represent the following standard concentrations 0.064, 0.32, 1.6, and 8 femtomoles μL^{-1}, respectively. Such concentrations were considered as the lower limit of detection for each synthetic peptide. They were determined based on a visual analysis of each peptide's transitions seeking for the identification of a clear alignment between all transitions for each standard peptide.



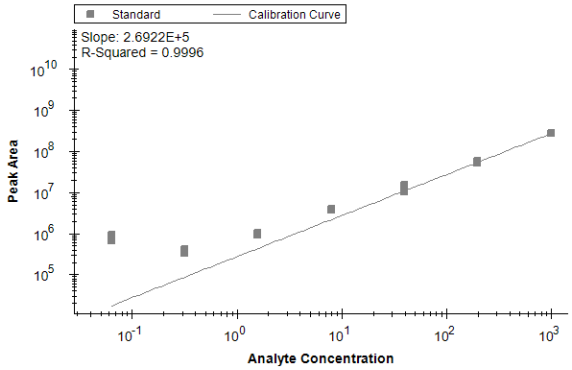
Calibration Curve: FVQAGSEVSALLGR



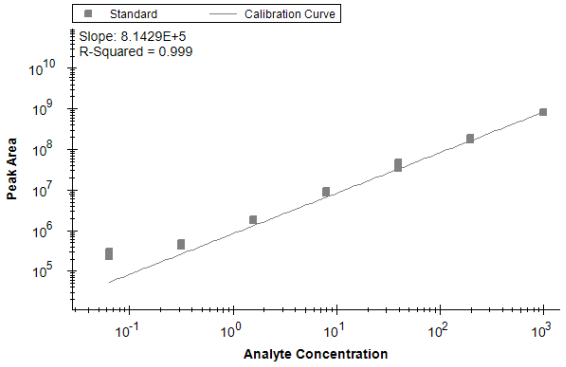
Calibration Curve: AVVAFTAVKPQLVVESPK



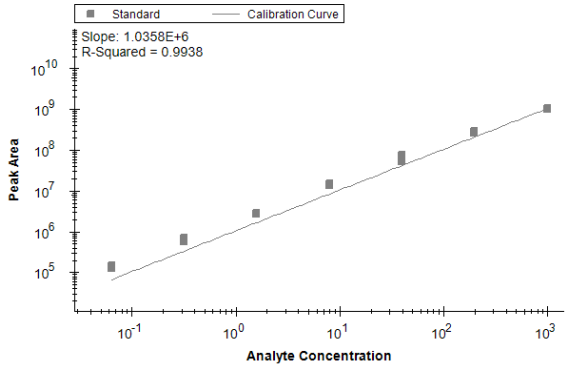
Calibration Curve: IFNVLGEPVDDLGPVDTR



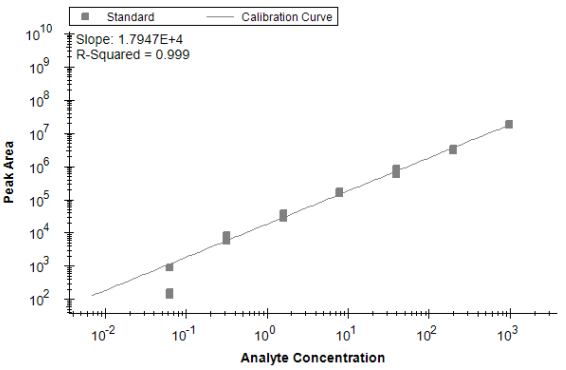
Calibration Curve: ELVNGELVAK



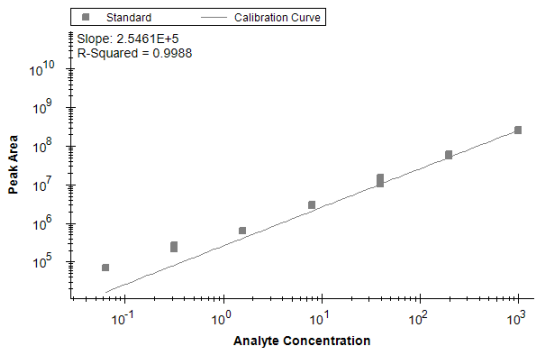
Calibration Curve: AAAEDAYFLVK



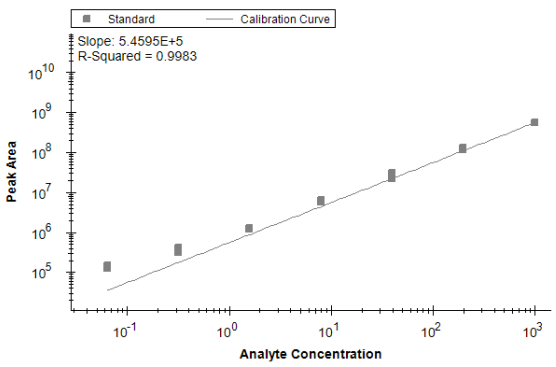
Calibration Curve: AFNPTQAEETYSMVTANR



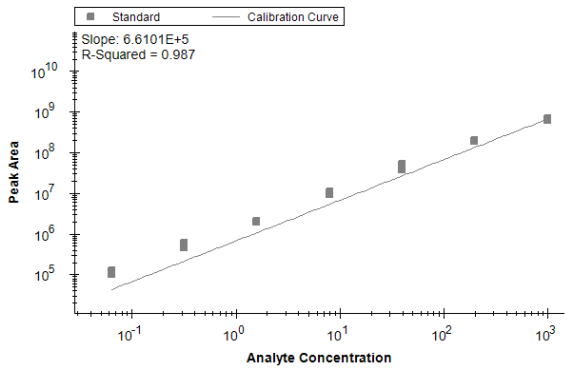
Calibration Curve: WDDGDTVFQLDR



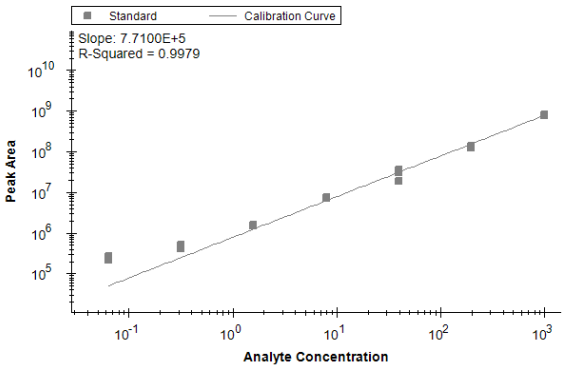
Calibration Curve: VTEVDKDTFWPIVNSAPDK



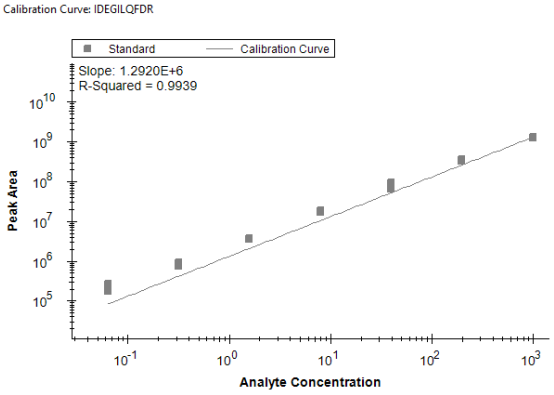
Calibration Curve: LYLDGAVLR



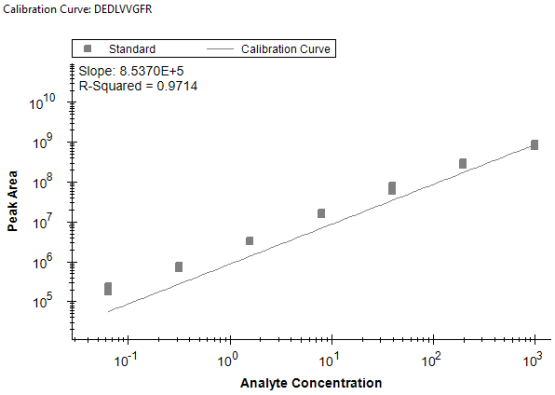
Calibration Curve: SEGEDPIGGSPARR



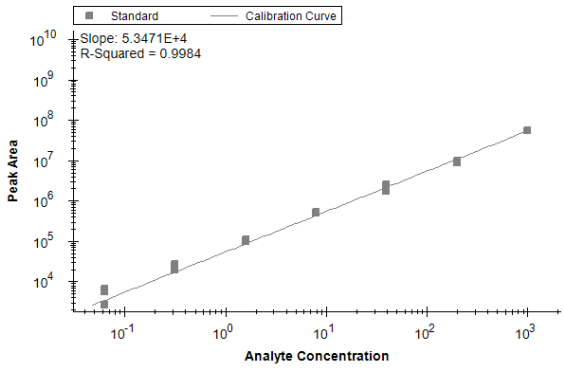
Calibration Curve: IDEGILQFDR



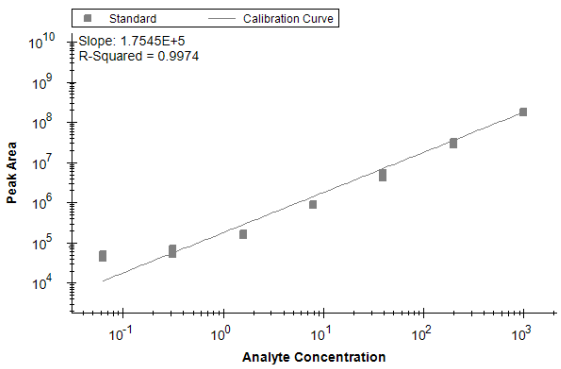
Calibration Curve: DEDLVVGFR



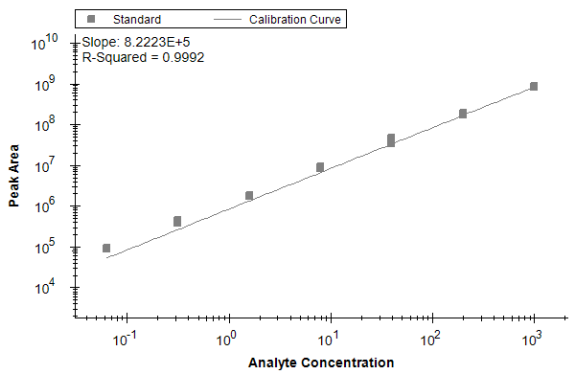
Calibration Curve: GPPPTASPPPGAPATSSSSR



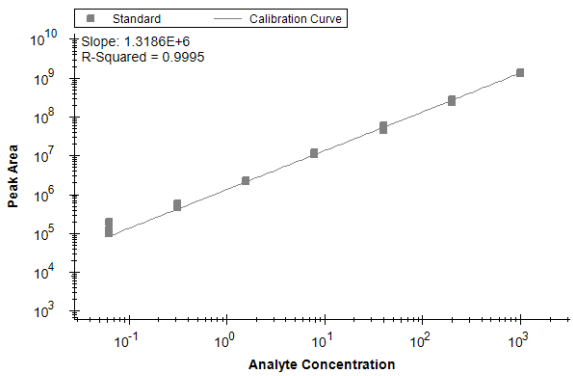
Calibration Curve: FWATVDVGRGGSLEFVGRSR



Calibration Curve: GLITSGENFQSK



Calibration Curve: LALSPLSR



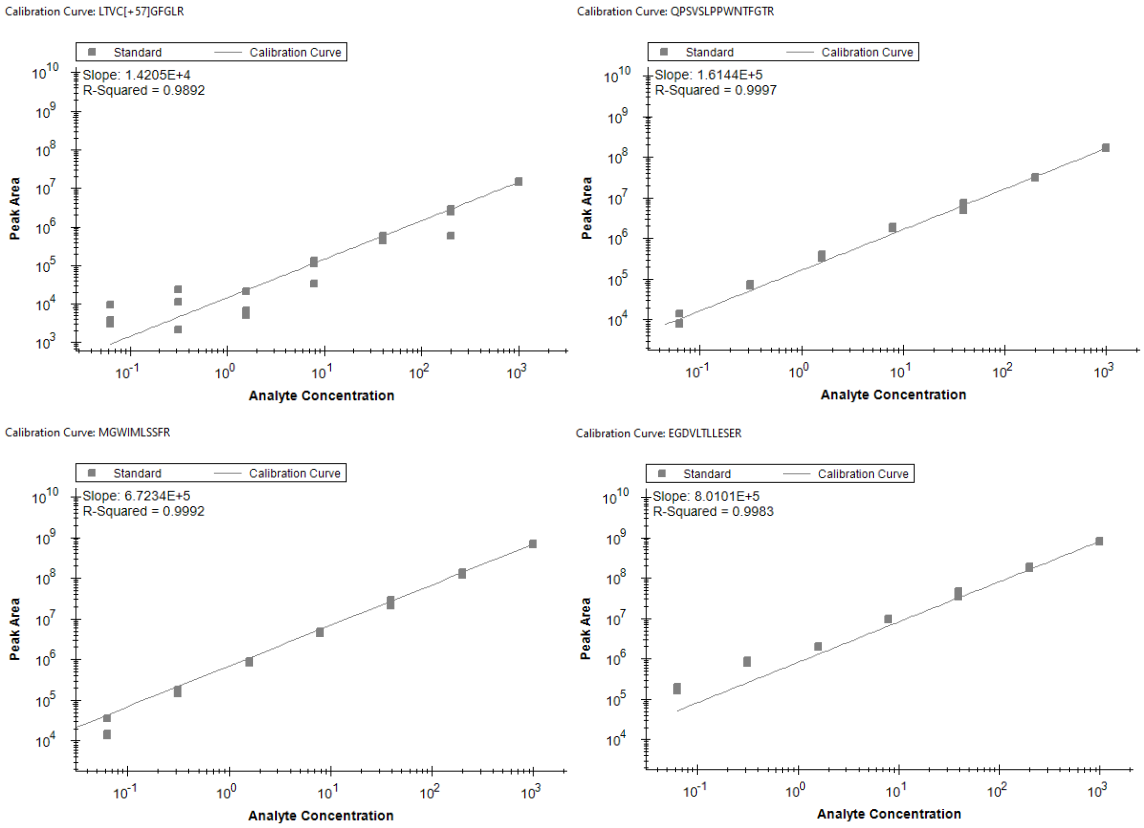


Figure S4 - Linear range response for each synthetic peptide defined by 7 concentrations at 1:4 dilution ratios (0.064 femtomoles μL^{-1} to 1 picomoles μL^{-1}). Three technical replicates were acquired at each concentration for each peptide. Linear regressions through zero were fit into each synthetic peptide. X (analyte concentration) and Y (peak area) axes were log transformed.

APPENDIX - Supplementary Tables

All supplementary tables can be downloaded upon email request to the following email: gabriellemes_13@hotmail.com or tiago.balbuena@unesp.br

Supplementary Table 1 - Accession number correspondence of novel identifications achieved by spectral correlations according to the *Eucalyptus grandis* BRASUZ1 genome annotation v2.0.

Supplementary Table 2 - List of all known proteins identified by LC-MS/MS in *Eucalyptus grandis* leaf samples, quantitative values, and summary of statistical analyses.

Supplementary Table 3 - List of all novel peptides identified by both spectral correlations and *de novo* peptide sequence in *Eucalyptus grandis* leaf samples, quantitative values, and summary of statistical analyses.

Supplementary Table 4 - Mapping assignment data used for novel peptides achieved by spectral correlations against *E. grandis* genome (Egrandis1_0 reference Annotation Release 101).

Supplementary Table 5 - Terms enriched by Fisher's exact test of *E. grandis* most abundant proteins over different seasons with and without water restriction imposition.

Supplementary Table 6 - List of differentially abundant known proteins in *E. grandis* leaf samples accomplished by pairwise comparisons (LFQ intensities of water-restricted over non-water-restricted values) of each season.

Supplementary Table 7 - Detailed description of differentially abundant novel peptides based on ANOVA analysis of *E. grandis* leaf samples under no water restriction treatment.

Supplementary Table 8 - Detailed description of differentially abundant novel peptides based on quantitative pairwise comparisons of water-restricted against non-water-restricted *E. grandis* leaf samples for each season.

Supplementary Table 9 - Evaluated parameters for protein structural analyses and protein template selection on t-test and ANOVA-significant peptides.

Supplementary Table 10 - Selected novel peptides for MRM targeted assays and previous results of DDA injections.

Supplementary Table 11 – Final acquisition method for MRM targeted assays.