11-2022

Occurrence of Hydroxyproline in Proteomes of Higher Plants

Olivia Huffman
University of Nebraska-Lincoln, ohuffman2@huskers.unl.edu

OCCURRENCE OF HYDROXYPROLINE IN PROTEOMES OF HIGHER PLANTS

by

Olivia K. Huffman

A THESIS

Presented to the Faculty of
The Graduate College at the University of Nebraska
In Partial Fulfillment of Requirements
For the Degree of Master of Science

Major: Food Science and Technology

Under the Supervision of Professors Joseph L. Baumert
and Philip Johnson

Lincoln, Nebraska

November 2022

OCCURRENCE OF HYDROXYPROLINE IN PROTEOMES OF HIGHER PLANTS

Olivia K. Huffman, M.S.

University of Nebraska, 2022

Advisors: Joseph L. Baumert and Philip Johnson

Food allergies affect millions of individuals across the United States and worldwide. Peanut allergies are among the most severe food allergies because of their potentially life-threatening symptoms and lifelong persistence. Potent peanut allergen, Ara h 2, is known to contain an amino acid motif containing the posttranslational modification, hydroxyproline (HyP). HyP is associated with immunogenic response when present both in Ara h 2 and in timothy grass pollen allergen, Phl p 1.

To further explore the presence of HyP in higher plants and specifically to investigate its potential presence in commonly allergenic plants, a study of 26 plant seeds was conducted using hydrolyzed amino acid analysis (HAA) and data-dependent acquisition (DDA) through liquid chromatography and tandem mass spectrometry (LC-MS/MS). Curated protein databases allowed for database searches using PEAKS software. Samples for which no database could be procured were analyzed using *de novo* sequencing.

Results showed detection of HyP in 25 out of 26 plant seed samples. HyP sites were classified into one of four tiers based on the quality of the database used to identify a given site. To further refine the identified HyP sites and to increase confidence in their position and identity, a manual analysis approach was performed in addition to software analyses. This approach was successful in reducing the number of sites for each sample as well as increasing the confidence of those sites. Peanut presented as a clear outlier in

the number of HyP sites in software and *de novo* analyses both before and after refinement by manual analysis.

The results indicate that species across Viridiplantae possess the machinery to perform prolyl hydroxylation. Furthermore, these data indicate that peanut is unique in its quantity of HyP sites even when normalized to the total number of proline residues.

ACKNOWLEDGEMENTS

This thesis and all the research contributing to it was completed purely by the power of my savior, Jesus. He has provided for my every need – physical, mental, spiritual, and otherwise – since setting foot in Lincoln. During times I didn’t think I would survive, He was there to hold me up. During times that have been so sweet, He was there to thank. All glory to God, the King of Kings!

Dr. Lee, I couldn’t have done this without you. Thank you for all your help with technical skills and an even bigger thank you for being my friend. I appreciate every moment you took to help me along and for teaching me how to be a scientist.

Dr. Justin Marsh, even though you are the daintiest man I have ever met, you are a great biochemist. Thank you for teaching me all I know about lab techniques and experimental design and for buying me all those cups of coffee.

Julie Nordlee, thank you from the bottom of my heart for loving and supporting me during the good times and hard times. Your wisdom and care have left their mark on my heart forever.

Thank you to my advisors, Dr. Baumert and Dr. Johnson, for giving me this opportunity and for supporting me along the way. I have loved growing and developing as part of your lab and am grateful for how it has shaped me.

Thank you to my lab mates, both past and present, for making FARRP such a special place to work. Liyun, I have thoroughly enjoyed sitting next to you these past couple of years – your presence alone brightens our [dark] office. Emily, I can’t tell you how grateful I am for your friendship. This graduate program was challenging and you have been a breath of fresh air during times when I felt stale. Sara, thank you for all you have been to me here in Lincoln. I am grateful to have walked this journey with you and I am better for knowing you. Jess, you inspire me daily to be a better version of myself. Thank you for reminding me not to take things too seriously and most of all, thank you for your kindness. Thank you also to my lab mates Maggie, Niloofar, Tengfei, Ellenor, Jenna,

Muhammad, Dr. Shyamali Jayasena, Dr. Bini Ramachadran and Dr. Shimin Chen for helping me to develop as a scientist and as a person.

Big Riss, I'm so thankful for your friendship, support, and advice and for creating a space to let loose and have fun. Thank you for encouraging my renewed love for reading and for trash tv. You're the peanut butter to my jelly.

To my Huskers, thank you for being my recess. I love you and can't thank you enough for accepting me into the family. I am eternally grateful for the chance to be a part of the Red Team and for letting me have a small part in continuing to build the most special program in the country. 2022 B1G Tourney Champs → 2023 WCWS. GBR!

Finally, thank you to my family. The past couple of years have been a time of growth and new beginnings, and I wouldn't have survived without you supporting me every step of the way. Thank you for loving me selflessly and generously – you are more than I ever could have asked for. Thank you also to Arlo for being the best emotional supporter on planet Earth. I love you guys.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: POST TRANSLATIONAL MODIFICATIONS RELEVANT TO ALLERGIES AND THE STUDY OF HYDROXYPROLINE IN PLANT PROTEINS BY LABEL-FREE MASS SPECTROMETRY

1.1 Food allergy

In the United States, it is estimated that 10% of individuals suffer from food allergy[1,2] and the rate appears to be increasing[3-5], though rates vary depending on assessment by self-reporting, skin-prick test (SPT), or by the gold standard for food allergy diagnosis, a double-blind placebo- controlled food challenge (DBPCFC)[6].

One common characteristic shared amongst many allergenic food sources is that the respective food's most abundant proteins also represent the food's major allergenic proteins[7-9]. In legumes and tree nuts for example, this trend can be observed as seed storage proteins (SSPs) make up most of the allergenic proteins in seeds such as peanut, soybean, almonds, cashews, and green peas[10].

Processing of allergenic foods has the potential to alter conformational epitopes by affecting the secondary, tertiary and/or quaternary structure of the protein, thus leading to reduced IgE binding to those epitopes. For example, heating the milk protein, β-lactoglobulin, to 70˚C – 80˚C can result in protein denaturation and unfolding, while heating over 100˚C can produce protein aggregation which functions to mask conformational epitopes and may ultimately decrease IgE reactivity and allergenicity[11]. In plants, the prolamin superfamily includes SSPs which are particularly thermostable due to cysteine residues and subsequent disulfide bond formation, and which retain their digestion-resistant, allergenic properties even after processing[12,13]. While processing may affect conformational epitopes to varying degrees, linear epitopes are likely to remain which will retain the potential to interact with the immune system.

In peanuts, processing is known to influence allergenicity. For example, in a study of sensitization to peanut using a mouse model, boiled peanut showed reduced sensitization compared to roasted peanuts. Furthermore, roasted peanuts elicited an increased allergic reaction compared to raw or boiled peanuts[14,15]. One theory suggests that this decrease in allergenicity was caused by the observed transfer of low-molecular-weight proteins into the water during cooking[14] which would include 2S albumin and potent peanut allergen, Ara h 2[16]. This resulted in a lower proportion of these allergens in the final food product, the boiled peanut. Reduction in 2S albumin allergen abundance was presumably responsible for the substantial decrease in observed allergenic responses.

Peanut allergy is a condition which affects approximately 2% of Americans[17,18] and can cause symptoms ranging from hives to anaphylaxis. Unlike milk and egg allergies, peanut allergies typically persist throughout a lifetime.

1.2 Peanut origins and growing patterns

The cultivated peanut, *Arachis hypogeae*, is widely utilized for oil, peanut butter, and snack consumption in the United States and worldwide with 50,310,111 metric tons of peanuts produced in 2021[19]. Its nutritive properties and relatively low production expenses make peanut a popular, low-cost food around the world.

The genus *Arachis*, including *A. hypogea,* commonly known as peanut or groundnut, originated in South America, likely in southwestern Brazil or northeast Paraguay[20] and belongs to the legume family *Fabaceae*. The peanut cultivated in present-day is an allotetraploid that evolved from the hybridization event of two diploid species, *A. duranesis* and *A. ipaensis*[21-24]. The allergenic protein Ara h 2.02 is highly homologous to Ara i 2 with both containing the motif DPYSpS three times[25]. Similarly, Ara h 2.01 is highly homologous to Ara d 2. Grabiele, et al.[26] suggests there may have been another tetraploid, *Arachis monticola*, which served as an intermediate between the diploid ancestors and gave rise to today's cultivated peanut, *A. hypogaea*.

Arachis members are unique from others in the Fabaceae family because of their geocarpic development patterns, defined as the process of underground (subterranean) fruit production; peanuts present flowers above-ground, containing the main organs for self-fertilization. After fertilization, the underground ovary matures into fruit and pod enlargement occurs[27,28]. Geocarpy is relatively uncommon but can be observed in some species in both tropical and arid climates[29]. Both climates have hostile elements – in arid climates, moisture is scarce and in tropical climates the soil can become highly weathered[30]. These environments may yield advantages to geocarpic adaptation; if a mother plant matures in a microsite containing sufficient moisture or nutrients in an otherwise scarce environment, it likely also makes a good environment for its seeds. The climate of the regions of Brazil and Paraguay associated with the diploid ancestors of the cultivated peanut are tropical, consistent with an environment that may encourage geocarpic adaptation patterns.

There is limited literature available on geocarpic growth; however, in a 2005 review of African and Madigascan Flora, refining of the term by division into three subcategories of active, passive, and geophytic geocarpy was suggested[29]. In active geocarpy, the flower blooms above ground and the plant works to actively bury the seed, a strategy most associated with peanut (*Arachis hypogeae)*, but also occurs in *Trifolium suterraneum*, *Vigna subterranean*, and *Macrotyloma geocarpum,* all of which interestingly belong to the Leguminosae family[29,31]. Passive geocarpy refers to the burying of the seeds by natural events (i.e. wind, water) rather than actively by the plant, and geophytic geocarpy refers to instances where the flower grows aboveground while the ovary, which becomes the fruit, stay below ground for the entirety of development[29].

This counter-intuitive method of seed dispersal is typically observed in environments with low nutrient density because the cluster effect of underground seed dispersal limits the distance seeds can travel, thus increasing probability that seeds land in favorable growing conditions. If the mother plant can reach reproductive maturity in that environment, it is likely its seedlings would also have greater chance of survival than in surrounding areas; therefore, it would be advantageous to allow them to germinate with close proximity to where they are produced[32].

There are four main peanut market types consumed in the United States and other western countries: Runner, Spanish, Valencia, and Virginia. All four market types are highly comparable in protein content at ~23% - 26% protein[33], protein profile, allergen content[34] and allergenicity[35].

1.3 Immunoglobulin E (IgE)

Immunoglobulin E (IgE) is one of five human immunoglobulins along with IgA, IgG, IgD, and IgM[36]. Like other monomeric antibodies, it is made up of two heavy chains and two light chains held together by two disulfide bonds. The C-terminal end of the protein contains the constant domain, a section of the heavy chain which determines the class and isotype of the antibody. The Fc region occurs here and allows for binding to cell surfaces, either-to FcεRI or FcεRII receptors. The N-termini of both the light and heavy chains, referred to as the F(ab) region, make up the variable domain and include a conformation that is complementary to a specific individual antigen. IgE, along with IgM, retains an additional constant heavy domain (CH) for a total of four, compared to the other immunoglobulin classes which maintain only three CH domains. While other classes exist primarily as pentamers (IgM) or dimers (IgA), IgE exists solely as a monomer. Plasma levels of IgE tend to be lower than any of the other classes of immunoglobulins, and 10,000-fold to 50,000-fold lower than serum levels of IgG[37,38]. Uniquely, half of the IgE content of the body can be found in tissues and bound to mast cells rather than in serum alone[39].

1.4 IgE-mediated hypersensitivity (allergy)

Development of an allergy occurs in two stages. The first stage is known as sensitization during which a foreign pathogenic protein is recognized and an immune response commences. T-helper cells (T_h) are activated and class-switching is initiated, a process in which plasma B-cells change from IgM production to antigen-specific IgE production. These antibodies have a high affinity for Fc_{ε} receptors and thus, readily bind to the receptors on mast cell or basophil surfaces. The binding of antibodies to the cell surface of these granulocytes completes the process of sensitization. The receptor at the cell surface is stabilized by IgE binding, and human basophil FcεRI expression is directly correlated with serum IgE levels[40].

After sensitization, there is no immediate response from the body. However, if a subsequent exposure to that same antigen does occur, the second stage of allergy development occurs. During the effector phase, the antigen binds to two cytotropic antibodies in a process known as "cross-linking" to initiate degranulation of the mast cell or basophil. Degranulation releases mediators into the surrounding tissue including histamine and other vasoactive molecules[41]. These pro-inflammatory molecules can cause symptoms typically associated with an allergic reaction including vasodilation, increased vascular permeability, fluid accumulation, and swelling.

The site on the antigen to which antibodies bind is of particular importance. This region, referred to as the "epitope" of the antigen, is responsible for the binding that initiates degranulation and triggers the allergic response and is therefore the subject of much research. Epitope binding is unique to, and dependent on, the structure and chemical makeup of the region of the antigen specific to the antibody. Epitopes can be classified as either linear or conformational. Linear epitopes are dictated by the primary structure of the protein and are determined by the properties of the specific amino acid sequence. Conversely, conformational epitopes comprise of a combination of amino acids and modifications when folded and can therefore be specific to primary, secondary, tertiary, and quaternary protein structures.

Interestingly, IgE-specific sensitization to an allergen, food allergens included, is not sufficient for elicitation of an allergic reaction[42]. In a UK population of 933 children, 11.8% of individuals were considered sensitized to peanut and only 22.4% of sensitized individuals possessed peanut allergy confirmed by OFC or DBPCFC[43]. This is concurrent with several studies indicating greater instances of sensitization to peanut compared to peanut allergy[44,45]. It is also notable that measurements of peanut sensitization only provide a limited prediction of peanut allergy severity[46]. There are several possible explanations for this phenomenon. If a patient has IgE specific to only a single epitope on an allergen, it would be unable to perform IgE crosslinking on a granulocyte, thus resulting in no degranulation. On the other hand, the IgE Fab region could have a low affinity for the allergen to which it binds. Though Fab regions are known to increase in binding affinity over subsequent exposures, it is unlikely that IgE binding affinity would be so low that it would fail to elicit a response. Furthermore, low IgE binding would be reflected in an an IgE binding assay. Finally, the most plausible explanation is that there is a high volume of IgG4 which is specific to the same antigen as IgE. IgG binding outcompetes that of IgE and therefore mitigates a possible allergenic response.

1.5 Collagen and cell wall protein characteristics

Understanding proteins is essential to understanding allergens. The collagen family of proteins is highly abundant in organisms across the animal kingdom and contains a high amount of an amino acid of particular interest, hydroxyproline (HyP). Collagen molecules are comprised of three coiled helices stabilized by hydrogen bonds[47,48]. The first high-resolution crystal structure of a triple-helical collagen-related peptide was presented in 1994[49]. Collagen has long been studied because of its uniqueness among other proteins; in the primary structure of this triple-helical molecule, glycine comprises every third residue (Gly-Xaa-Yaa-Gly, etc.)[47,48,50-52]. Furthermore, there is a high occurrence of proline and hydroxyproline in the other two positions (Xaa and Yaa)[47,48,50-52]. In collagen, hydroxylation does not occur by incorporation of hydroxylated free amino acids, but rather by modification of peptide-bound residues[53]. It is catalyzed after translation in the endoplasmic reticulum (ER) where folding of proteins occurs[53] . When the α1 chain of collagen IV (murine) was analyzed *in silico*, 322 total proline residues were found within the COL4A1 sequence[54]. Of these, only 54 proline residues were found in the Xaa position[54]. Proline residues in the Xaa position are subject to hydroxylation at their tertiary carbon[54] by collagen prolyl 3-hydroxylase (c-P3H), a relatively rare occurrence. Conversely, of the 322 total proline residues found in the COL4A1 sequence, 213 occurred in the Yaa position where they are subject to hydroxylation at their quaternary carbon[54] by collagen prolyl 4-hydroxylase (c-P4H). In collagen type IV, 50—60% of all proline residues are hydroxylated[55,56].

The motifs Gly-X-HyP and Gly-Pro-Y are relevant to specific aspects of the structure's stability: proline residues contribute stiffness and rigidity to the molecule while HyP residues contribute elasticity and flexibility[57]. Amide hydrogen atoms function to stabilize α and β secondary structure. Peptide bonds in proline residues lack these hydrogen bonds allowing prolines to be utilized for certain structural features (e.g. tight turns) and to increase protein backbone rigidity in globular proteins[58]. Furthermore, intermolecular hydrogen bonding is provided by 4-HyP at the Yaa position of the Gly-

Xaa-Yaa motif, allowing for stabilization of the triple helix and maturation of the molecule.

In addition to its abundance in collagen, HyP is also readily found in plant cell walls. Hydroxyproline-rich glycoproteins (HRGPs) are a class of modified proteins which have undergone either N-glycosylation or O-glycosylation. In plants, low-oxygen status is sensed by the Cys branch of the N-degron pathway[59,60]. Degradation signals, determined by N-terminal residues of cellular proteins, are referred to as N-degrons[59,61,62]. The identity of a protein's N-terminal residue can be related to its *in vivo* half-life, a phenomenon known as the N-end rule[61,62].

1.6 Allergenic relevance of hydroxyproline

While the roles of HyP in collagen and HRGPs are well-established, the function of the modified proline in soluble cell proteins is less understood. Two of the major peanut allergens, Ara h 2 and Ara h 6, are seed storage proteins classified as 2S-albumins. These conglutins belong to the prolamin superfamily and have 59% sequence homology[63,64]. 2S albumins have a stabilized core and conserved disulfide bridges, making them resistant to proteolytic digestion and a significant candidate to elicit systemic allergic reactions. Five α-helices and four disulfide bonds comprise Ara h 2, a protein which is recognized by over 90% of peanut-allergic individuals and the most potent peanut allergen[65,66]. Two major isoforms of Ara h 2 have been identified as Ara h 2.01 and 2.02, which contain insertions of 14 and 26 amino acids, respectively, which are not contained in Ara h 6. It was previously thought that Ara h 2 did not contain post-translational modifications apart from disulfide bridges[63]. It is now well understood that these insertions form a flexible surface loop on which a specific motif, DPYSpS, is repeated two (Ara h 2.01) or three (Ara h 2.02) times; these regions contain post-translationally hydroxylated prolines in the second proline in the motif[67,68]. These hydroxyproline-containing motifs are known to be responsible for optimal IgE binding (approximately 50% of IgE binding on Ara h 2) and their absence have been shown to decrease mediator release in humanized rat basophil leukemia (RBL) cell mediator release assays, a more biologically relevant assessment of the potency of allergens compared to IgE binding assays alone[63,68,69].

Additionally, even when the motifs are present but are produced recombinantly in *E. coli*, the IgE binding is significantly lower than that to proteins produced by a eukaryotic organism, *Nicotiana benthamiana*, or the native peanut protein[64,68,70]. It is now acknowledged that bacteria lack the machinery to perform proline hydroxylation[71]; therefore, while the authentic amino acid sequence can be produced in a prokaryotic expression system, no HyP modifications occur as they do in eukaryotic expression.

In Timothy grass pollen, one linear epitope has been identified which covers the N-terminus of allergen Phl p 1[72]. In this 10-residue epitope, there are three proline residues and two were found to be hydroxylated and contribute to IgE-binding of the epitope[72,73].

1.7 IgE function and the hygiene hypothesis

While IgE is associated with allergic response, it is widely hypothesized that it was once a mechanism to protect humans from harmful toxins or parasites. This phenomenon can be described by the hygiene hypothesis, referring to the lack of exposure to pathogens in an ultra-clean environment which may hinder immune system development and render the system vulnerable to recognizing non-harmful antigens as pathogens and initiating an inflammatory response[74]. Some hypothesize that the symptoms commonly associated with allergy (e.g., scratching, vomiting, coughing, diarrhea, sneezing) are derived from mechanisms for expelling parasites too large to be endocytosed by traditional immune mechanisms. This hypothesis is also supported by the immunosuppressive effects observed during helminth infection[75-78]. Another idea suggests that the evolution of the allergic response developed to defend against immediate danger caused by toxins. The "toxin hypothesis" cites the aforementioned symptoms instead as plausible mechanisms for expelling toxins, as well as a drop in blood pressure caused by histamine which could slow the rate at which an allergen circulates in the bloodstream[79]. While there is some debate as to whether the "original" function of IgE was to protect against long-lived parasitic worms (helminths), but it is indisputable that reactions to these macroparasites share many parallels to allergic reactions as they elicit a Th2-type response which leads to antibody class-switching and IgE production[80].

Though reduced in recent decades, rates of soil-transmitted helminths (STH) are still endemic in many tropical and sub-tropical regions[81]. In urban, westernized societies where STH infections are not endemic, allergies have higher reported prevalence compared to lower, albeit increasing, prevalence in tropical and subtropical populations[74,81]. It is important to note that climate is understood to play little, if any, role in allergy development and relies more significantly on social and economic environments. Human immunological reactions to both allergy and helminth infection elicit Th2-associated responses, and because of their similarity, allergic responses are thought to arise from a misdirected immune response to these parasitic worms[82].

Helminths rely on a collagen-rich exoskeleton referred to as the “cuticle” to mediate body function and interaction with the environment[53]. Mature collagen synthesized using a high amount of amino acids glycine, proline, and hydroxyproline provides structure and flexibility to this outer coating. Helminth infections trigger IL-10 release and, because of their anti-inflammatory properties, pose significant potential for therapeutic use[75-77]. IL-10 is known primarily as a suppressive cytokine and functions to decrease inflammatory response by inhibiting the release of cell mediators including histamine[83,84]. This role is critical for protecting the host against detrimental tissue damage from excessive mediator release[76-78].

While allergic response to collagen from many mammalian species is limited, IgE sensitization and cross-linking to fish collagen demonstrate that fish collagen is an important allergen in some fish-allergic individuals (including 50% of Japanese patients with fish allergy having detectable IgE to fish collagen)[85-88].

1.8 Mass spectrometry (MS) as a tool for post-translational modification (PTM) study

While researchers have uncovered many isoforms, sequences, cross-reactivities, and risk factors in recent decades, there is still much to be discovered regarding food allergies. Specifically, there is a gap in understanding the role that modifications play in IgE binding and elicitation of basophil and mast cell degranulation. It is known that post-translational modifications (PTMs) occur after translation and are not explicitly written into the DNA sequence. PTMs impact the chemical composition of their respective residue(s); therefore, they have the potential to impact both linear and conformational epitopes by increasing or decreasing affinity for the appropriate antibody. Additionally, PTMs can alter interactions between protein domains, thus adjusting the folding patterns and final tertiary and quaternary structure of a protein. PTMs can be grouped into two major classes: (1) modifications associated with structured regions which contribute to catalytic function, enzyme activity, or structure stabilization, and (2) modifications associated with disordered regions of proteins which rely on a low-affinity, high-specificity enzyme-substrate interaction[89]. Currently, the most reliable and specific means of studying PTMs is by using mass spectrometry (MS).

PEAKS 8.5 (Bioinformatics Solutions, Inc.) is a premier software tailored to analysis of proteomic discovery data from mass spectrometry. It performs data conversion, peptide and protein identification, PTM characterization, and result validation, all from the raw MS data file[90] (Figure 1.1).

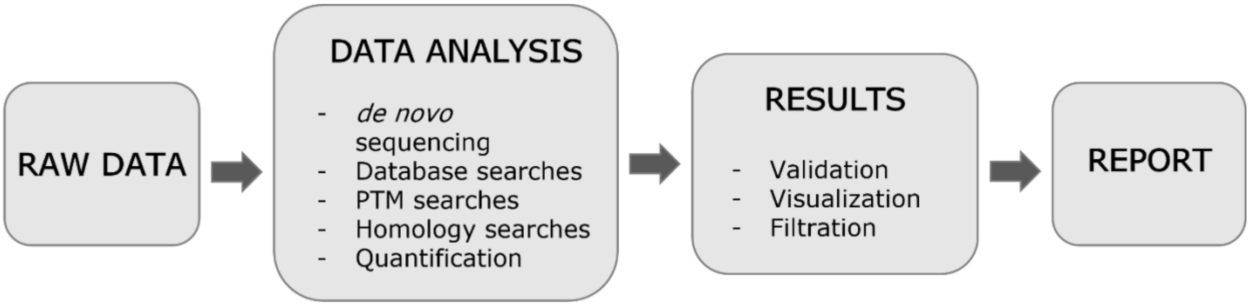


Figure 1.1: Proteomic data analysis workflow and capabilities

Raw data files can be uploaded to PEAKS 8.5 software which can perform *de novo* sequencing, database searches, posttranslational modification identification, homology searches, and quantification using internal standards. It then displays the data on a user-friendly interface for visualization, filtration, and validation. This figure was adapted from the PEAKS 8.5 User Manual[90].

1.8.1 Software parameters and mechanisms

To increase reliability of identification, the software deploys an automatic “decoy-fusion” method which releases a set number of decoy hits as part of the dataset, measured by false discovery rate (FDR). This metric can be adjusted to filter the data to a higher or lower degree of confidence.

The software also produces interactive visual graphics to clearly display peptide coverage, PTMs, and spectral abundance. Unfortunately, although PEAKS is confirmed to be superior as “demonstrated by publications and third-party evaluations,” there are some gaps in its ability to identify and evaluate PTMs in the manner pertinent to our research discussed in the following chapters of this thesis. For example, PEAKS quantifies PTMs and filters according to AScore or minimum ion intensity threshold.

The minimum ion intensity is the relative intensity that needs to be met or exceeded by a fragment containing the modification for the modification to be regarded as confident[90]. A MS2 spectrum must be observed at this relative intensity for the PTM to be considered confident. AScore, an alternative filter for PTMs, refers to the confidence of a PTM at a specific position in a peptide. Another useful PTM tool is the PTM Profiling application, which calculates the abundance of modified peptides versus unmodified peptides according to your confidence parameters.

In addition to confidence parameters and interactive graphics, features can be detected, deconvoluted, and refined using an expectation-maximization based algorithm to perform Label Free Quantification (LFQ)[90]. This is a semi-quantitative method in which MS1 spectra from multiple samples are matched by feature, aligned by retention time, and normalized to selected proteins. Because the m/z variation from peptide to peptide cannot be measured without internal heavy-isotope labeling, this method is only semi-quantitative and provides a relative comparison of intensities for matching features between replicates. Alternatively, quantification can be loosely predicted by spectral counting, a method in which the number of spectra for a given protein is identified[91].

The principle of heavy isotope labeling is that the heavy peptides have identical chemical properties to that of the native peptide, so it elutes off the column at the same time as the native peptide. Its m/z, on the other hand, will be different because the mass of the molecule is heavier. The ratio of the heavy-to-light peptide intensities can then be calculated for absolute quantification.

1.8.2 Electrospray MS

Electrospray ionization (ESI) is a method of ionization of a liquid sample and transition of the liquid sample to the gas phase. Droplets of the sample are dispersed as a highly ionized fine mist and passed down a pressure gradient where ions from the surface of the charged droplets are ejected into the gaseous phase. These ions are detected by the mass analyzer again where their molecular mass and ion intensity are recorded as precursor ions. Further detail can be obtained when precursor ions are fragmented, detected again, and analyzed by a second mass analyzer producing MS2 spectra[92].

1.8.3 Fragmentation

There are several methods of fragmentation to obtain MS2 spectra. One commonly used method is collision-induced dissociation (CID). This method utilizes an electrical potential to activate and accelerate protonated peptides and then allows them to collide with neutral gas molecules, inducing bond breakage and thus, fragmentation of the peptides into b- and y- ions. These fragment ions are detected by the mass analyzer and can be integrated using software to match them to their precursor ion. This allows overlapping of fragment ions to determine monoisotopic mass, and therefore identity, of individual residues or groups of residues within the precursor ion.

1.8.4 Label-free quantification (LFQ)

While there is a positive correlation between signal intensity and ion concentration in electrospray ionization (ESI) techniques[93], there is often a nonlinear relationship between the two variables because of variable ionization efficiency due to matrix interactions and ion chemistry[94,95].

Signal intensities can be influenced by a variety of factors during ESI including analyte size, ionization efficiency, etc. These variables between analytes in a complex mixture make it difficult to quantify analytes according to a single internal standard without the use of heavy isotope labeling. Some ions will fly differently than others. Therefore, features (otherwise known as detected precursor ions) can only be compared to each other when they detect the same ion across different samples. LFQ does this by identifying identical m/z ratios and aligning their retention times, thus overlaying the peak areas for comparison.

This is more limiting than the absolute quantification possible with heavy peptides; however, this is a cost- and time-effective alternative that can still provide information of relative abundance to comparable mass events.

1.9 Concluding remarks

Food allergies appear to be increasing worldwide, especially in westernized countries. Some hypothesize that the lack of traditional immune challenges, such as those associated with parasitic infections which are mitigated by modern-day hygiene practices, may be leading to the increased prevalence of food allergies. HyP is found on parasite proteins, timothy grass pollen allergen, Phl p 1, as well as the major peanut allergens, Ara h 2 and 6, which has led some to question whether HyP found on plant-based food proteins may increase the potential for allergic sensitization and elicitation. We set out to explore this potential association with HyP further by analyzing various seed storage proteins from foods using amino acid analyses and mass spectrometry techniques as detailed further in this thesis.

REFERENCES

(1) Gupta, R. S.; Warren, C. M.; Smith, B. M.; Jiang, J.; Blumenstock, J. A.; Davis, M. M.; Schleimer, R. P.; Nadeau, K. C. Prevalence and Severity of Food Allergies Among US Adults. *JAMA network open* **2019,** *2* (1), e185630.

(2) "Allergies and Hay Fever," Centers for Disease Control and Prevention, 2018.

(3) Nwaru, B. I.; Hickstein, L.; Panesar, S. S.; Roberts, G.; Muraro, A.; Sheikh, A.; Allergy, t. E. F.; Group, A. G. Prevalence of common food allergies in Europe: a systematic review and meta-analysis. **2014,** *69* (8), 992.

(4) Sicherer, S. H.; Sampson, H. A. Food allergy: A review and update on epidemiology, pathogenesis, diagnosis, prevention, and management. *Journal of Allergy and Clinical Immunology* **2018,** *141* (1), 41.

(5) Sicherer, S. H.; Muñoz-Furlong, A.; Godbold, J. H.; Sampson, H. A. US prevalence of self-reported peanut, tree nut, and sesame allergy: 11-year follow-up. *The Journal of allergy and clinical immunology* **2010,** *125* (6), 1322.

(6) Sicherer, S. H.; Warren, C. M.; Dant, C.; Gupta, R. S.; Nadeau, K. C. Food Allergy from Infancy Through Adulthood. *The Journal of Allergy and Clinical Immunology: In Practice* **2020,** *8* (6), 1854.

(7) Bannon, G. A. What makes a food protein an allergen? *Current Allergy and Asthma Reports* **2004,** *4* (1), 43.

(8) Shin, D. S.; Compadre, C. M.; Maleki, S. J.; Kopper, R. A.; Sampson, H.; Huang, S. K.; Burks, A. W.; Bannon, G. A. Biochemical and structural analysis of the IgE binding sites on ara h1, an abundant and highly allergenic peanut protein. *The Journal of biological chemistry* **1998,** *273* (22), 13753.

(9) Costa, J.; Villa, C.; Verhoeckx, K.; Cirkovic-Velickovic, T.; Schrama, D.; Roncada, P.; Rodrigues, P. M.; Piras, C.; Martín-Pedraza, L.; Monaci, L.et al. Are Physicochemical Properties Shaping the Allergenic Potency of Animal Allergens? *Clinical Reviews in Allergy & Immunology* **2022,** *62* (1), 1.

(10) Breiteneder, H.; Ebner, C. Molecular and biochemical classification of plant-derived food allergens. *The Journal of allergy and clinical immunology* **2000,** *106* (1 Pt 1), 27.

(11) Monaci, L.; Pilolli, R.; De Angelis, E.; Crespo, J. F.; Novak, N.; Cabanillas, B. In *Advances in Food and Nutrition Research*; Toldrá, F., Ed.; Academic Press, 2020; Vol. 93.

(12) Wickham, M.; Faulks, R.; Mills, C. In vitro digestion methods for assessing the effect of food structure on allergen breakdown. **2009,** *53* (8), 952.

(13) Bennetau-Pelissero, C., 2018, DOI:10.1007/978-3-319-54528-8_3-1 10.1007/978-3-319-54528-8_3-1.

(14) Vissers, Y. M.; Iwan, M.; Adel-Patient, K.; Stahl Skov, P.; Rigby, N. M.; Johnson, P. E.; Mandrup Müller, P.; Przybylski-Nicaise, L.; Schaap, M.; Ruinemans-Koerts, J.et al. Effect of roasting on the allergenicity of major peanut allergens Ara h 1 and Ara h 2/6: the necessity of degranulation assays. **2011,** *41* (11), 1631.

(15) Zhang, T.; Shi, Y.; Zhao, Y.; Tang, G.; Niu, B.; Chen, Q. Boiling and roasting treatment affecting the peanut allergenicity. *Annals of translational medicine* **2018,** *6* (18), 357.

(16) Mondoulet, L.; Paty, E.; Drumare, M. F.; Ah-Leung, S.; Scheinmann, P.; Willemot, R. M.; Wal, J. M.; Bernard, H. Influence of Thermal Processing on the Allergenicity of Peanut Proteins. *Journal of Agricultural and Food Chemistry* **2005,** *53* (11), 4547.

(17) Warren, C.; Lei, D.; Sicherer, S.; Schleimer, R.; Gupta, R. Prevalence and characteristics of peanut allergy in US adults. *Journal of Allergy and Clinical Immunology* **2021,** *147* (6), 2263.

(18) Lieberman, J. A.; Gupta, R. S.; Knibb, R. C.; Haselkorn, T.; Tilles, S.; Mack, D. P.; Pouessel, G. The global burden of illness of peanut allergy: A comprehensive literature review. *Allergy* **2021,** *76* (5), 1367.

(19) USDA - Foreign Agricultural Service, 2022.

(20) Simpson, C. E.; Krapovickas, A.; Valls, J. F. M. History of Arachis Including Evidence of A. hypogaea L. Progenitors. *Peanut Science* **2001,** *28* (2), 78.

(21) Husted, L. Cytological Studies an the Peanut, <i>Arachis</i>. II

Chromosome number, morphology and behavior, and their application to the problem of the origin of the cultivated forms. *CYTOLOGIA* **1936,** *7* (3), 396.

(22) Kochert, G.; Stalker, H. T.; Gimenes, M.; Galgaro, L.; Lopes, C. R.; Moore, K. RFLP and Cytogenetic Evidence on the Origin and Evolution of Allotetraploid Domesticated Peanut, Arachis hypogaea (Leguminosae). *American Journal of Botany* **1996,** *83* (10), 1282.

(23) Moretzsohn, M. C.; Gouvea, E. G.; Inglis, P. W.; Leal-Bertioli, S. C.; Valls, J. F.; Bertioli, D. J. A study of the relationships of cultivated peanut (Arachis hypogaea) and its most closely related wild species using intron sequences and microsatellite markers. *Annals of botany* **2013,** *111* (1), 113.

(24) Bertioli, D. J.; Cannon, S. B.; Froenicke, L.; Huang, G.; Farmer, A. D.; Cannon, E. K. S.; Liu, X.; Gao, D.; Clevenger, J.; Dash, S.et al. The genome sequences of Arachis duranensis and Arachis ipaensis, the diploid ancestors of cultivated peanut. *Nature Genetics* **2016,** *48* (4), 438.

(25) UniProt; European Bioinformatics Institute, 2022.

(26) Grabiele, M.; Chalup, L.; Robledo, G. A.; Seijo, G. J. P. S.; Evolution. Genetic and geographic origin of domesticated peanut as evidenced by 5S rDNA and chloroplast DNA sequences. **2012,** *298*, 1151.

(27) Smith, B. W. Arachis hypogaea. Aerial Flower and Subterranean Fruit. *American Journal of Botany* **1950,** *37* (10), 802.

(28) Ozias-Akins, P.; Breiteneder, H. The functional biology of peanut allergens and possible links to their allergenicity. *Allergy* **2019,** *74* (5), 888.

(29) Barker, N. A review and survey of basicarpy, geocarpy, and amphicarpy in the African and Madagascan flora. *Annals of the Missouri Botanical Garden* **2005,** *92*, 445.

(30) Lathwell, D. J.; Grove, T. L. Soil-Plant Relationships in the Tropics. *Annual Review of Ecology and Systematics* **1986,** *17*, 1.

(31) Kaul, V.; Koul, A. K.; Sharma, M. C. The underground flower. *Current Science* **2000,** *78* (1), 39.

(32) Harrison, R. D.; Rønsted, N.; Xu, L.; Rasplus, J.-Y.; Cruaud, A. Evolution of Fruit Traits in Ficus Subgenus Sycomorus (Moraceae): To What Extent Do Frugivores Determine Seed Dispersal Mode? *PLOS ONE* **2012,** *7* (6), e38432.

(33) Campos-Mondragón, M. G.; Calderón De La Barca, A. M.; Durán-Prado, A.; Campos-Reyes, L. C.; Oliart-Ros, R. M.; Ortega-García, J.; Medina-Juárez, L. A.; Angulo, O. Nutritional composition of new Peanut (Arachis hypogaea L.) cultivars. *Grasas y Aceites* **2009,** *60* (2), 161.

(34) Marsh, J. T.; Palmer, L. K.; Koppelman, S. J.; Johnson, P. E. Determination of Allergen Levels, Isoforms, and Their Hydroxyproline Modifications Among Peanut Genotypes by Mass Spectrometry. **2022,** *3*.

(35) Koppelman, S. J.; Jayasena, S.; Luykx, D.; Schepens, E.; Apostolovic, D.; de Jong, G. A. H.; Isleib, T. G.; Nordlee, J.; Baumert, J.; Taylor, S. L.et al. Allergenicity attributes of different peanut market types. *Food and Chemical Toxicology* **2016,** *91*, 82.

(36) Ishizaka, K.; Ishizaka, T. Identification of γE-Antibodies as a Carrier of Reaginic Activity. **1967,** *99* (6), 1187.

(37) Garman, S. C.; Wurzburg, B. A.; Tarchevskaya, S. S.; Kinet, J. P.; Jardetzky, T. S. Structure of the Fc fragment of human IgE bound to its high-affinity receptor Fc epsilonRI alpha. *Nature* **2000,** *406* (6793), 259.

(38) Bloebaum, R. M.; Dharajiya, N.; Grant, J. A. Mechanisms of IgE-mediated allergic reactions. *Clinical allergy and immunology* **2004,** *18*, 65.

(39) Poole, J. A.; Rosenwasser, L. J. The role of immunoglobulin E and immune inflammation: implications in allergic rhinitis. *Curr Allergy Asthma Rep* **2005,** *5* (3), 252.

(40) Stone, K. D.; Prussin, C.; Metcalfe, D. D. IgE, mast cells, basophils, and eosinophils. *The Journal of allergy and clinical immunology* **2010,** *125* (2 Suppl 2), S73.

(41) Krystel-Whittemore, M.; Dileepan, K. N.; Wood, J. G. Mast Cell: A Multi-Functional Master Cell. **2016,** *6*.

(42) Santos, A. F.; Du Toit, G.; O'Rourke, C.; Becares, N.; Couto-Francisco, N.; Radulovic, S.; Khaleva, E.; Basting, M.; Harris, K. M.; Larson, D.et al. Biomarkers of severity and threshold of allergic reactions during oral peanut challenges. *Journal of Allergy and Clinical Immunology* **2020,** *146* (2), 344.

(43) Nicolaou, N.; Poorafshar, M.; Murray, C.; Simpson, A.; Winell, H.; Kerry, G.; Härlin, A.; Woodcock, A.; Ahlstedt, S.; Custovic, A. Allergy or tolerance in children sensitized to peanut: prevalence and differentiation using component-resolved diagnostics. *The Journal of allergy and clinical immunology* **2010,** *125* (1), 191.

(44) Ta, V.; Weldon, B.; Yu, G.; Humblet, O.; Neale-May, S.; Nadeau, K. Use of Specific IgE and Skin Prick Test to Determine Clinical Reaction Severity. *British journal of medicine and medical research* **2011,** *1* (4), 410.

(45) Wainstein, B. K.; Studdert, J.; Ziegler, M.; Ziegler, J. B. Prediction of anaphylaxis during peanut food challenge: usefulness of the peanut skin prick test (SPT) and specific IgE level. **2010,** *21* (4p1), 603.

(46) Datema, M. R.; Lyons, S. A.; Fernández-Rivas, M.; Ballmer-Weber, B.; Knulst, A. C.; Asero, R.; Barreales, L.; Belohlavkova, S.; de Blay, F.; Clausen, M.et al. Estimating the Risk of Severe Peanut Allergy Using Clinical Background and IgE Sensitization Profiles. **2021,** *2*.

(47) Ramachandran, G. N.; Kartha, G. Structure of collagen. *Nature* **1955,** *176* (4482), 593.

(48) Rich, A.; Crick, F. H. C. The Structure of Collagen. *Nature* **1955,** *176* (4489), 915.

(49) Bella, J.; Eaton, M.; Brodsky, B.; Berman, H. M. Crystal and molecular structure of a collagen-like peptide at 1.9 A resolution. *Science (New York, N.Y.)* **1994,** *266* (5182), 75.

(50) Schroeder, W. A.; Kay, L. M.; LeGette, J.; Honnen, L.; Green, F. C. The Constitution of Gelatin. Separation and Estimation of Peptides in Partial Hydrolysates. *Journal of the American Chemical Society* **1954,** *76* (13), 3556.

(51) Brodsky, B.; Ramshaw, J. A. The collagen triple-helix structure. *Matrix biology : journal of the International Society for Matrix Biology* **1997,** *15* (8-9), 545.

(52) Bella, J.; Brodsky, B.; Berman, H. M. Hydration structure of a collagen peptide. *Structure* **1995,** *3* (9), 893.

(53) Winter, A. D.; McCormack, G.; Myllyharju, J.; Page, A. P. Prolyl 4-hydroxlase activity is essential for development and cuticle formation in the human infective parasitic nematode Brugia malayi. *The Journal of biological chemistry* **2013,** *288* (3), 1750.

(54) Basak, T.; Vega-Montoto, L.; Zimmerman, L. J.; Tabb, D. L.; Hudson, B. G.; Vanacore, R. M. Comprehensive Characterization of Glycosylation and Hydroxylation of Basement Membrane Collagen IV by High-Resolution Mass Spectrometry. *Journal of proteome research* **2016,** *15* (1), 245.

(55) Kleinman, H. K.; McGarvey, M. L.; Liotta, L. A.; Robey, P. G.; Tryggvason, K.; Martin, G. R. Isolation and characterization of type IV procollagen, laminin, and heparan sulfate proteoglycan from the EHS sarcoma. *Biochemistry* **1982,** *21* (24), 6188.

(56) Kefalides, N. A. In *International Review of Connective Tissue Research*; Hall, D. A.;Jackson, D. S., Eds.; Elsevier, 1973; Vol. 6.

(57) Ghanaeian, A.; Soheilifard, R. Mechanical elasticity of proline-rich and hydroxyproline-rich collagen-like triple-helices studied using steered molecular dynamics. *Journal of the mechanical behavior of biomedical materials* **2018,** *86*, 105.

(58) Murrali, M. G.; Piai, A.; Bermel, W.; Felli, I. C.; Pierattelli, R. Proline Fingerprint in Intrinsically Disordered Proteins. **2018,** *19* (15), 1625.

(59) Licausi, F.; Kosmacz, M.; Weits, D. A.; Giuntoli, B.; Giorgi, F. M.; Voesenek, L. A.; Perata, P.; van Dongen, J. T. Oxygen sensing in plants is mediated by an N-end rule pathway for protein destabilization. *Nature* **2011,** *479* (7373), 419.

(60) Gibbs, D. J.; Holdsworth, M. J. Every Breath You Take: New Insights into Plant and Animal Oxygen Sensing. *Cell* **2020,** *180* (1), 22.

(61) Varshavsky, A. The N-end rule pathway and regulation by proteolysis. *Protein science : a publication of the Protein Society* **2011,** *20* (8), 1298.

(62) Varshavsky, A. N-degron and C-degron pathways of protein degradation. **2019,** *116* (2), 358.

(63) Lehmann, K.; Schweimer, K.; Reese, G.; Randow, S.; Suhr, M.; Becker, W. M.; Vieths, S.; Rösch, P. Structure and stability of 2S albumin-type peanut allergens: implications for the severity of peanut allergic reactions. *The Biochemical journal* **2006,** *395* (3), 463.

(64) Mueller, G. A.; Gosavi, R. A.; Pomés, A.; Wünschmann, S.; Moon, A. F.; London, R. E.; Pedersen, L. C. Ara h 2: crystal structure and IgE binding distinguish two subpopulations of peanut allergic patients by epitope diversity. *Allergy* **2011,** *66* (7), 878.

(65) Koppelman, S. J.; Wensing, M.; Ertmann, M.; Knulst, A. C.; Knol, E. F. Relevance of Ara h1, Ara h2 and Ara h3 in peanut-allergic patients, as determined by immunoglobulin E Western blotting, basophil–histamine release and intracutaneous testing: Ara h2 is the most important peanut allergen. **2004,** *34* (4), 583.

(66) Palmer, G. W.; Dibbern, D. A., Jr.; Burks, A. W.; Bannon, G. A.; Bock, S. A.; Porterfield, H. S.; McDermott, R. A.; Dreskin, S. C. Comparative potency of Ara h 1 and Ara h 2 in immunochemical and functional assays of allergenicity. *Clinical immunology (Orlando, Fla.)* **2005,** *115* (3), 302.

(67) Chatel, J. M.; Bernard, H.; Orson, F. M. Isolation and characterization of two complete Ara h 2 isoforms cDNA. *International archives of allergy and immunology* **2003,** *131* (1), 14.

(68) Bernard, H.; Guillon, B.; Drumare, M. F.; Paty, E.; Dreskin, S. C.; Wal, J. M.; Adel-Patient, K.; Hazebrouck, S. Allergenicity of peanut component Ara h 2: Contribution of conformational versus linear hydroxyproline-containing epitopes. *The Journal of allergy and clinical immunology* **2015,** *135* (5), 1267.

(69) Hazebrouck, S.; Guillon, B.; Paty, E.; Dreskin, S. C.; Adel-Patient, K.; Bernard, H. Variable IgE cross-reactivity between peanut 2S-albumins: The case for measuring IgE to both Ara h 2 and Ara h 6. *Clinical and Experimental Allergy* **2019,** *49* (8), 1107.

(70) Üzülmez, Ö.; Kalic, T.; Mayr, V.; Lengger, N.; Tscheppe, A.; Radauer, C.; Hafner, C.; Hemmer, W.; Breiteneder, H. The Major Peanut Allergen Ara h 2 Produced in Nicotiana benthamiana Contains Hydroxyprolines and Is a Viable Alternative to the E. Coli Product in Allergy Diagnosis. **2021,** *12*.

(71) Glenting, J.; Poulsen, L. K.; Kato, K.; Madsen, S. M.; Frøkiær, H.; Wendt, C.; Sørensen, H. W. Production of Recombinant Peanut Allergen Ara h 2 using Lactococcus lactis. *Microbial cell factories* **2007,** *6*, 28.

(72) Petersen, A.; Suck, R.; Hagen, S.; Cromwell, O.; Fiebig, H.; Becker, W.-M. Group 13 grass allergens: Structural variability between different grass species and analysis of proteolytic stability. *Journal of Allergy and Clinical Immunology* **2001,** *107* (5), 856.

(73) Petersen, A.; Becker, W. M.; Schlaak, M. Characterization of grass group I allergens in timothy grass pollen. *The Journal of allergy and clinical immunology* **1993,** *92* (6), 789.

(74) Bach, J.-F. The hygiene hypothesis in autoimmunity: the role of pathogens and commensals. *Nature Reviews Immunology* **2018,** *18* (2), 105.

(75) Hang, L.; Blum, A. M.; Setiawan, T.; Urban, J. P., Jr.; Stoyanoff, K. M.; Weinstock, J. V. Heligmosomoides polygyrus bakeri infection activates colonic Foxp3+ T cells

enhancing their capacity to prevent colitis. *Journal of immunology (Baltimore, Md. : 1950)* **2013,** *191* (4), 1927.

(76) van der Vlugt, L. E.; Labuda, L. A.; Ozir-Fazalalikhan, A.; Lievers, E.; Gloudemans, A. K.; Liu, K. Y.; Barr, T. A.; Sparwasser, T.; Boon, L.; Ngoa, U. A.et al. Schistosomes induce regulatory features in human and mouse CD1d(hi) B cells: inhibition of allergic inflammation by IL-10 and regulatory T cells. *PLoS One* **2012,** *7* (2), e30883.

(77) Redpath, S. A.; Fonseca, N. M.; Perona-Wright, G. Protection and pathology during parasite infection: IL-10 strikes the balance. **2014,** *36* (6), 233.

(78) Suzuki, K.; Yamada, M.; Kurakake, S.; Okamura, N.; Yamaya, K.; Liu, Q.; Kudoh, S.; Kowatari, K.; Nakaji, S.; Sugawara, K. Circulating cytokines and hormones with immunosuppressive but neutrophil-priming potentials rise after endurance exercise in humans. *European journal of applied physiology* **2000,** *81* (4), 281.

(79) Profet, M. The Function of Allergy: Immunological Defense Against Toxins. *The Quarterly Review of Biology* **1991,** *66* (1), 23.

(80) Fitzsimmons, C.; Falcone, F.; Dunne, D. Helminth Allergens, Parasite-Specific IgE, and Its Protective Role in Human Immunity. **2014,** *5*.

(81) Pullan, R. L.; Smith, J. L.; Jasrasaria, R.; Brooker, S. J. Global numbers of infection and disease burden of soil transmitted helminth infections in 2010. *Parasites & Vectors* **2014,** *7* (1), 37.

(82) Artis, D.; Maizels, R. M.; Finkelman, F. D. Allergy challenged. *Nature* **2012,** *484* (7395), 458.

(83) Webster, H. C.; Gamino, V.; Andrusaite, A. T.; Ridgewell, O. J.; McCowan, J.; Shergold, A. L.; Heieis, G. A.; Milling, S. W. F.; Maizels, R. M.; Perona-Wright, G. Tissue-based IL-10 signalling in helminth infection limits IFNγ expression and promotes the intestinal Th2 response. *Mucosal Immunology* **2022**, DOI:10.1038/s41385-022-00513-y 10.1038/s41385-022-00513-y.

(84) Iyer, S. S.; Cheng, G. Role of interleukin 10 transcriptional regulation in inflammation and autoimmune disease. *Critical reviews in immunology* **2012,** *32* (1), 23.

(85) Kobayashi, Y.; Akiyama, H.; Huge, J.; Kubota, H.; Chikazawa, S.; Satoh, T.; Miyake, T.; Uhara, H.; Okuyama, R.; Nakagawara, R.et al. Fish collagen is an important panallergen in the Japanese population. **2016,** *71* (5), 720.

(86) Hamada, Y.; Nagashima, Y.; Shiomi, K. Identification of collagen as a new fish allergen. *Bioscience, biotechnology, and biochemistry* **2001,** *65* (2), 285.

(87) Kuehn, A.; Hilger, C.; Hentges, F. Anaphylaxis provoked by ingestion of marshmallows containing fish gelatin. *Journal of Allergy and Clinical Immunology* **2009,** *123* (3), 708.

(88) Kalic, T.; Kamath, S. D.; Ruethers, T.; Taki, A. C.; Nugraha, R.; Le, T. T. K.; Humeniuk, P.; Williamson, N. A.; Hira, D.; Rolland, J. M.et al. Collagen—An Important Fish Allergen for Improved Diagnosis. *The Journal of Allergy and Clinical Immunology: In Practice* **2020,** *8* (9), 3084.

(89) Xie, H.; Vucetic, S.; Iakoucheva, L. M.; Oldfield, C. J.; Dunker, A. K.; Obradovic, Z.; Uversky, V. N. Functional anthology of intrinsic disorder. 3. Ligands, post-translational modifications, and diseases associated with intrinsically disordered proteins. *Journal of proteome research* **2007,** *6* (5), 1917.

(90) Bioinformatics Solutions, I. Waterloo, ON, Canada, 2017.

(91) Zou, Z.; Tao, T.; Li, H.; Zhu, X. mTOR signaling pathway and mTOR inhibitors in cancer: progress and challenges. *Cell & Bioscience* **2020,** *10* (1), 31.

(92) Ho, C. S.; Lam, C. W.; Chan, M. H.; Cheung, R. C.; Law, L. K.; Lit, L. C.; Ng, K. F.; Suen, M. W.; Tai, H. L. Electrospray ionisation mass spectrometry: principles and clinical applications. *The Clinical biochemist. Reviews* **2003,** *24* (1), 3.

(93) Voyksner, R. D.; Lee, H. Investigating the use of an octupole ion guide for ion storage and high-pass mass filtering to improve the quantitative performance of electrospray ion trap mass spectrometry. **1999,** *13* (14), 1427.

(94) Trufelli, H.; Palma, P.; Famiglini, G.; Cappiello, A. An overview of matrix effects in liquid chromatography-mass spectrometry. *Mass spectrometry reviews* **2011,** *30* (3), 491.

(95) Jeanne Dit Fouque, D.; Maroto, A.; Memboeuf, A. Internal Standard Quantification Using Tandem Mass Spectrometry of a Tryptic Peptide in the Presence of an Isobaric Interference. *Analytical Chemistry* **2018,** *90* (24), 14126.

(96) Peters, R. L.; Koplin, J. J.; Gurrin, L. C.; Dharmage, S. C.; Wake, M.; Ponsonby, A.-L.; Tang, M. L. K.; Lowe, A. J.; Matheson, M.; Dwyer, T.et al. The prevalence of food allergy and other allergic diseases in early childhood in a population-based study: HealthNuts age 4-year follow-up. *Journal of Allergy and Clinical Immunology* **2017,** *140* (1), 145.

(97) Skolnick, H. S.; Conover-Walker, M. K.; Koerner, C. B.; Sampson, H. A.; Burks, W.; Wood, R. A. The natural history of peanut allergy. *Journal of Allergy and Clinical Immunology* **2001,** *107* (2), 367.

(98) De Leon, M. P.; Glaspole, I. N.; Drew, A. C.; Rolland, J. M.; O'Hehir, R. E.; Suphioglu, C. Immunological analysis of allergenic cross-reactivity between peanut and tree nuts. *Clinical & Experimental Allergy* **2003,** *33* (9), 1273.

(99) Migueres, M.; Dávila, I.; Frati, F.; Azpeitia, A.; Jeanpetit, Y.; Lhéritier-Barrand, M.; Incorvaia, C.; Ciprandi, G. Types of sensitization to aeroallergens: definitions, prevalences and impact on the diagnosis and treatment of allergic respiratory disease. *Clinical and translational allergy* **2014,** *4*, 16.

(100) Du Toit, G.; Katz, Y.; Sasieni, P.; Mesher, D.; Maleki, S. J.; Fisher, H. R.; Fox, A. T.; Turcanu, V.; Amir, T.; Zadik-Mnuhin, G.et al. Early consumption of peanuts in infancy is associated with a low prevalence of peanut allergy. *Journal of Allergy and Clinical Immunology* **2008,** *122* (5), 984.

(101) Lack, G. Epidemiologic risks for food allergy. *Journal of Allergy and Clinical Immunology* **2008,** *121* (6), 1331.

(102) Goodman RE, E. M., Ferreira F, Sampson HA, van Ree R, Vieths S, Baumert JL, Bohle B, Lalithambika S, Wise J, Taylor SL; Mol. Nutr. Food Res, 2016.

(103) In *Public Law 117-11*; Congress, t., Ed., 2021.

(104) In *21 USC 301 note*, 2004.

(105) Bublin, M.; Breiteneder, H. Cross-reactivity of peanut allergens. *Curr Allergy Asthma Rep* **2014,** *14* (4), 426.

(106) Mennini, M.; Dahdah, L.; Mazzina, O.; Fiocchi, A. Lupin and Other Potentially Cross-Reactive Allergens in Peanut Allergy. *Current Allergy and Asthma Reports* **2016,** *16* (12), 84.

(107) Ajilogba, C. F.; Olanrewaju, O. S.; Babalola, O. O. Improving Bambara Groundnut Production: Insight Into the Role of Omics and Beneficial Bacteria. **2022,** *13*.

(108) Tan, X. L.; Azam-Ali, S.; Goh, E. V.; Mustafa, M.; Chai, H. H.; Ho, W. K.; Mayes, S.; Mabhaudhi, T.; Azam-Ali, S.; Massawe, F. Bambara Groundnut: An Underutilized Leguminous Crop for Global Food Security and Nutrition. *Frontiers in nutrition* **2020,** *7*, 601496.

(109) Shimizu, M.; Igasaki, T.; Yamada, M.; Yuasa, K.; Hasegawa, J.; Kato, T.; Tsukagoshi, H.; Nakamura, K.; Fukuda, H.; Matsuoka, K. Experimental determination of proline hydroxylation and hydroxyproline arabinogalactosylation motifs in secretory proteins. **2005,** *42* (6), 877.

(110) Petersen, A.; Schramm, G.; Schlaak, M.; Becker, W. M. Post-translational modifications influence IgE reactivity to the major allergen Phl p 1 of timothy grass pollen. *Clinical and experimental allergy : journal of the British Society for Allergy and Clinical Immunology* **1998,** *28* (3), 315.

(111) Camerini, S.; Mauri, P. The role of protein and peptide separation before mass spectrometry analysis in clinical proteomics. *Journal of Chromatography A* **2015,** *1381*, 1.

(112) Gorres, K. L.; Raines, R. T. Prolyl 4-hydroxylase. *Critical Reviews in Biochemistry and Molecular Biology* **2010,** *45* (2), 106.

(113) Han, X.; He, L.; Xin, L.; Shan, B.; Ma, B. PeaksPTM: Mass spectrometry-based identification of peptides with unspecified modifications. *Journal of proteome research* **2011,** *10* (7), 2930.

(114) Stanke, M.; Steinkamp, R.; Waack, S.; Morgenstern, B. AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic acids research* **2004,** *32* (Web Server issue), W309.

(115) Stanke, M.; Keller, O.; Gunduz, I.; Hayes, A.; Waack, S.; Morgenstern, B. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic acids research* **2006,** *34* (Web Server issue), W435.

(116) Chen, S.; Downs, M. L. Proteomic Analysis of Oil-Roasted Cashews Using a Customized Allergen-Focused Protein Database. *Journal of proteome research* **2022,** *21* (7), 1694.

(117) FAIRsharing.org: HWG, 2016, DOI: 10.25504/FAIRsharing.srgkaf 10.25504/FAIRsharing.srgkaf.

(118) Geiselhart, S.; Hoffmann-Sommergruber, K.; Bublin, M. Tree nut allergens. *Molecular Immunology* **2018,** *100*, 71.

(119) Goetz, D. W.; Whisman, B. A.; Goetz, A. D. Cross-reactivity among edible nuts: double immunodiffusion, crossed immunoelectrophoresis, and human specific IgE serologic surveys. *Annals of Allergy, Asthma & Immunology* **2005,** *95* (1), 45.

(120) Andorf, S.; Borres, M. P.; Block, W.; Tupa, D.; Bollyky, J. B.; Sampath, V.; Elizur, A.; Lidholm, J.; Jones, J. E.; Galli, S. J.et al. Association of Clinical Reactivity with Sensitization to Allergen Components in Multifood-Allergic Children. *The Journal of Allergy and Clinical Immunology: In Practice* **2017,** *5* (5), 1325.

(121) Asero, R.; Mistrello, G.; Roncarolo, D.; Amato, S. Walnut-induced anaphylaxis with cross-reactivity to hazelnut and Brazil nut. *Journal of Allergy and Clinical Immunology* **2004,** *113* (2), 358.

(122) Duan, G.; Walther, D. The Roles of Post-translational Modifications in the Context of Protein Interaction Networks. *PLOS Computational Biology* **2015,** *11* (2), e1004049.

(123) STULEMEIJER, I. J. E.; JOOSTEN, M. H. A. J. Post-translational modification of host proteins in pathogen-triggered defence signalling in plants. **2008,** *9* (4), 545.

(124) Bond, A. E.; Row, P. E.; Dudley, E. Post-translation modification of proteins; methodologies and applications in plant sciences. *Phytochemistry* **2011,** *72* (10), 975.

(125) Dai, Z.; Hooker, B. S.; Quesenberry, R. D.; Thomas, S. R. Optimization of Acidothermus cellulolyticus Endoglucanase (E1) Production in Transgenic Tobacco Plants by Transcriptional, Post-transcription and Post-translational Modification. *Transgenic Research* **2005,** *14* (5), 627.

(126) Xiao, Y.; Vecchi, M. M.; Wen, D. Distinguishing between Leucine and Isoleucine by Integrated LC–MS Analysis Using an Orbitrap Fusion Mass Spectrometer. *Analytical Chemistry* **2016,** *88* (21), 10757.

CHAPTER 2: PROTEIN EXTRACTION AND ANALYSIS OF AMINO ACID COMPOSITION

2.1 Introduction

Food allergies affect approximately 5.6% of individuals under the age of 18 in the United States and up to 10.8% of Americans overall[1,2]. Peanut allergies, which are estimated to affect just under 2% of Americans[1,96], can be among the most severe of food allergies with potential life-threatening symptoms and lifelong persistence[97]. The sera of over 90% of peanut-allergic individuals binds with protein Ara h 2 which is known to be post translationally modified on a significant linear epitope for IgE binding[65,66,68,70]. This region is located on a surface of a flexible loop and contains either two (Ara h 2.01) or three (Ara h 2.02) consistently identifiable hydroxyproline (HyP) sites[67,68]. Though HyP appears to be an important aspect in IgE binding to peanut protein, a critical step in elicitation of an allergic reaction, relatively little is known about the occurrence or abundance of HyP in other consumed plants.

To examine whether the presence of HyP within seed storage proteins of known allergenic sources correlates to increase opportunity for allergenicity, we included a wide variety of seed plants within a diverse portion of a diverse portion of the clade, Viridiplantae, that have histories of consumption in this study. With this approach, known food allergens from all different plant families as well as plants related to known food allergens could be examined for HyP presence. Some known allergenic tree nuts include walnut, pecan, cashew, almond, and Brazil nut. Some known allergenic legumes include soybean, peanut, green pea, and lupin bean. These seeds were included as well as related plants that are commonly consumed or increasing in consumption. Additionally, wheat and sesame were included because of their known allergen status. Collectively, grasses, sedges, trees, and legumes were included. The relationships between sample species can be seen in Figure 1. In all, 26 plant seed samples were analyzed. Of these, only *Pinus edulis* belongs to Acrogymnospermae. Two species, *Triticum aesetivum* and *Cyperus esculentus,* are classified as monocotyledons and both belong to the order Poales; however, *T. aesetivum* belongs to the grass family while *C. esculentus* belongs to the sedge family. One sample, *Macadamia integrifolia* belongs to the Protea family. All other samples are defined as eudicotyledons, with species coming from the orders Sapinales, Fabales, Rosales, Fagales, Lameales, and Ericales.

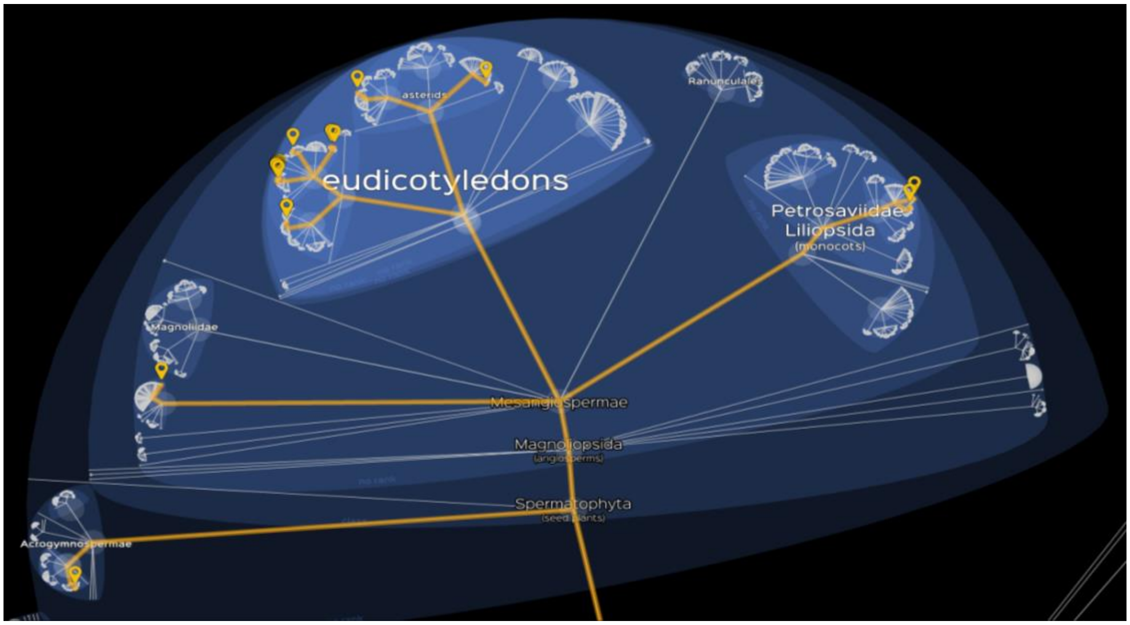


Figure 2.1: Relationship among all samples studied across Viridiplantae

Source: Lifemap, NCBI

This chapter aims to assess a wide variety of plant seed samples with known consumption in the human population for their extraction efficiencies and HyP presence using 2D Quantification, SDS-PAGE, and amino acid analysis. Emphasis is placed on exploring the relationship between proline and HyP abundance.

2.2 Hypothesis

Part 1:

Null hypothesis (H_0): There is no correlation between HyP and total proline concentrations in ground samples as determined by HAA.

Part 2:

Null hypothesis (H_0): There is no correlation between HyP and total proline concentrations in extracted samples as determined by HAA.

2.3 Materials and Methods

2.3.1 Sample Preparation

2.3.1.1 Sourcing

Sample selection was based on the following criteria:

- **Classification as a legume or tree nut**: Because peanut is known to contain HyP sites on an important immunogenic antigen, many taxonomically related species (i.e. other members of the legume family) were included in the study. Furthermore, tree nuts were included because of the potential for cross-reactivity of peanut-allergic individuals to some tree-nut allergens[98]. Worth mentioning is that co-sensitization may occur as well as or instead of cross-reactivity Co-sensitization occurs when structurally different IgE molecules are present and bind allergens simultaneously[99], meaning that an individual could develop a separate tree nut allergy in addition to a peanut allergy and the two could happen simultaneously. This can present difficulty in determining which process is occurring.

- **Consumption**: Consumption frequency was also considered when defining the sample pool for this study. Plants which are either commonly consumed or increasing in consumption are of particular importance because environmental exposure to certain foods creates the potential for sensitization to those foods. Furthermore, in societies where some foods are more frequently consumed after infancy but not during weaning, higher instances of allergy to those foods could occur[100,101]. For these reasons, foods with common or increasing consumption rates were prioritized.
- **Known allergenicity**: Plants which have known allergen status according to AllergenOnline[102] were prioritized over plants without known IgE reactivity or allergenicity.
- **Presence in the Big Nine**: All major plant groups included in the Big Nine were included in the study regardless of plant family[103,104].
- **Proteome availability**: Species for which comprehensive protein databases were available were prioritized over those without databases. While all samples of known allergenicity were included regardless of database availability, some tree nuts and legumes were selected because of high sequence abundance. This allowed high quality data to be

acquired on tree nut and legume samples to provide valuable insight into the amino acid profiles of these species.

Each commercially available sample was sourced raw or in dried form wherever possible to avoid the potential for processing induced effects on HyP content. When possible, the seed (shell not included) was ordered. For walnut, the seed was separated from other plant material before grinding. It is likely that, though product packaging stated “raw, unprocessed,” many of the tree nuts may have undergone a blanching process before commercial sale.

2.3.1.2 Homogenization

Each sample was ground by hand using mortar and pestle and liquid nitrogen to achieve a fine powder. In total, approximately 5-10g of each sample was ground and immediately stored at -80C.

2.3.1.3 Extraction

All extraction protocols used 18-ohm water.

After homogenization, each sample was extracted in triplicate in a reducing and denaturing buffer of 6M urea (BioRad), 2M thiourea (Sigma-Aldrich), 20 mM dithiothreitol (DTT) (Sigma-Aldrich), 50 mM Tris (Sigma-Aldrich) pH 8.8 buffer at 50 mg/mL (w/v) so as not to saturate the extraction buffer. Approximately 500 mg of sample was measured into a 15 mL tube with the appropriate amount of buffer to achieve a 50 mg/mL concentration. Each sample was well-mixed (30s vortex) to ensure the sample was sufficiently dispersed in the extraction buffer to increase the opportunity for protein solubilization. The samples were then heated in a heating, shaking water bath set to 60˚C at 200 rpm for 10 minutes, mixed for ~15 seconds, for 10 minutes at room temperature (RT), briefly mixed again and returned to the heating, shaking water bath for 10 additional minutes. Upon removal from the water bath, samples were centrifuged at 3000xg for 10 minutes at room temperature (RT). 1 mL of supernatant was transferred into two separate 1.5 mL microcentrifuge tubes (500 µL each) and centrifuged for 10 minutes at 17000xg at RT for additional clarification. The supernatant (total 850 µL) was then transferred to a single 2 mL Eppendorf Safe-Lock® microcentrifuge tube and thoroughly mixed. This pool was then aliquoted into four tubes of 400 µL each for storage to minimize freeze/thaw of samples and stored immediately at -20˚C until needed.

2.3.1.4 Sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE)

Sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) gels (NuPAGE™, 4-12%) were run under reducing conditions at 200V constant for 35 minutes (BioRad) to confirm presence of proteins indicating successful extraction. One biological replicate of each sample was loaded onto the gel at 7 µg of protein per well with the exception of tiger nut which was loaded at 1.8 µg of protein. The tiger nut extract had an extraordinarily low soluble protein concentration compared to the other samples and the volume required to achieve 7 µg of protein would have exceeded the loading capacity of the wells of the gels. After electrophoretic separation, the gels were then fixed using 50% methanol and 10% glacial acetic acid. Bands were visualized using

Coomassie brilliant blue-R250 staining solution (BioRad) followed by de-staining and imaging.

2.3.1.5 2D Quantification

All 2D quant protocols used 18-ohm water.

The soluble protein concentration in each extraction was quantified using a 2D Quantification (2DQ) kit (Cytiva). The 2DQ relies on copper binding capacity to proteins. It measures the unbound copper concentration meaning that a higher absorbance reading indicates a higher amount of unbound copper, correlating with a lower amount of protein present. Conversely, a lower optical density indicates low amounts of unbound copper which indicates high amounts of protein binding and a high protein concentration. Half assays were performed to determine protein quantities between 5 μg - 25 μg.

2.3.1.6 Hydrolyzed amino acid analysis (HAA)

To assess the amino acid profiles of ground and extracted samples, hydrolyzed amino acid analyses (HAA) were performed by the Proteomics & Metabolomics Facility, Nebraksa Center for Biotechnology at the University of Nebraska-Lincoln. The following methods come from the official analysis reports:

“For HAA of ground samples, a small aliquot of ~ 5 mg was used for the acidic hydrolysis in 0.5% phenol/6N HCl after oxidation of sulfur-containing amino acids.

Before hydrolysis, an oxidation step using performic acid was done to convert Cys and Met into Cysteic acid (CyA) and Methionine Sulfone (MetS). The hydrolysates were dried down and the final pellets were dissolved in 20 mM HCl and derivatized using the AccQ-Tag reagent from Waters. The concentrations in µmoles/g and in mg/g DW for each amino acid detected in each sample are found in the excel file attached with the summary report (Ground_HAA_RawData.xls). The concentrations are calculated using a series of standards dilution run before the samples. Note: Cys and Met are oxidized into Cya and MetSO2. Gln and Asn are converted to Glu and Asp, respectively. Trp is destroyed during hydrolysis.

“For HAA of extracted samples, a test sample and a ubiquitin standard were run in the urea buffer. The results showed that the urea/Tris buffer is interfering with the assay. Another test was done using ubiquitin standard and performing a protein precipitation. 50 ug were precipitated with 100% acetone and incubated overnight at -20C. The pellet was washed one more time and used for the hydrolysis. The results comparing to a ubiquitin standard not precipitated showed that the precipitation was not responsible for protein loss and that the urea/buffer was efficiently removed so that it does not interfere with the derivatization. An aliquot of 50 ug of proteins (based on the protein assay provided) was used for acetone precipitation prior to acidic hydrolysis in 0.5% phenol/6N HCl after oxidation of sulfur-containing amino acids. Before hydrolysis, an oxidation step using performic acid was done to convert Cys and Met into Cysteic acid (CyA) and Methionine Sulfone (MetS). The hydrolysates were dried down and the final pellets were dissolved in 20 mM HCl and derivatized using the AccQ-Tag reagent from Waters.

“The concentrations in pmoles/uL and in ng/uL for each amino acid detected in each sample are found in the excel file attached with the summary report (Extracted_HAA_RawData.xls). The concentrations are calculated using a series of standards dilution run before the samples. Note: Cys and Met are oxidized into Cya and MetSO2. Gln and Asn are converted to Glu and Asp, respectively. Trp is destroyed during hydrolysis.

“The UPLC for both analyses was a 1290 Agilent Infinity II. The column used was ACCQ-TAG Ultra C18 1.7µm, 2.1x100 mm, with mobile phase A: 100% Eluent A, mobile phase B: 10:90 Eluent B: Milli-Q water, mobile phase C: Milli-Q water, and mobile phase D: 100% Eluent B. The flow rate was 0.7 mL/min, the column oven was at 48˚C. The runtime and gradient can be viewed in the supplementary document (Extract_HAA_OfficialReport).”

2.3.2 Statistical analysis

To assess correlation between multiple variables across MS data exported through PEAKS, Spearman r ranked correlation tests were performed using GraphPad Prism 9.1.0 for Windows, GraphPad Software, San Diego, California, USA, www.graphpad.com. Additionally, ROUT outlier tests were performed using GraphPad Prism 9.1.0 for Windows, GraphPad Software, San Diego, California, USA, www.graphpad.com.

2.4 Results & Discussion

2.4.1 Sample selection

As previously stated, samples were selected according to classification as a legume or tree nut, level of consumption (both in the US and abroad), known allergenicity, presence in the Big Nine, and proteome availability.

Because many of the previous studies involving hydroxyproline modification of allergens were performed using peanut proteins, peanut is of particular interest. Peanuts are part of the legume family and known to be both commonly allergenic and a potent allergenic food source. To investigate related organisms for similar patterns, peanut and other members of the legume family were included. In addition, several tree nuts were included because of the potential occurrence of IgE cross reactivity between peanut and tree nut allergens[105].

To increase relevance of the study, seeds and nuts that are frequently consumed or increasing in consumption were included. Sensitization to lupin bean, a legume related to peanut, has been observed in 15-20% of known peanut-allergic individuals[106]. Lupin beans are not a commonly consumed food in the US; however, they are used widely across Europe as a pickled snack food and as a protein source in bakery items and vegan meats. Peanuts are consumed with higher frequency in North America and lower frequency across Europe where hazelnuts are more popular. The Bambara groundnut is gaining traction as a sustainable leguminous crop in semi-arid regions of Africa because of its resilience in drought and marginalized soils as well as its well-rounded nutritive properties[107,108]. It has potential to be used for improvement of food and nutrient security[108], and thus, is expected to increase in consumption in coming decades. Because of this forecasted increase, it is relevant to include in this exploratory study of the molecular makeup of legumes and potentially allergenic foods.

In addition to classification as a legume or tree nut and consumption levels, known allergenicity and presence in the Big Nine were also considered. For feasibility of the study, known allergen status assisted in selecting specific legumes or tree nuts of interest. For example, almond and hazelnut are established food allergens[102] while chestnut is not; thus, almond and hazelnut were included. Wheat and sesame were included because they are two of the nine major allergens required to be explicitly labeled when contained in foods in the US[103,104].

Table 2.1: Sample information

Common name	**Scientific name**	**Database size (# sequences)**	**Source**
Almond (nonpareil)	*Prunus dulcis*	32104	FARRP
Bambara groundnut	*Vigna subterranea*	108636	Etsy
Baru nut	*Dipteryx alata*	225	Whole Foods
Brazil nut	*Bertholletia excelsa*	2497	FARRP
Cashew	*Anacardium occidentale*	130789	FARRP
Chickpea	*Cicer arietinum*	30798	Hy-Vee
Cowpea	*Vigna unguiculata*	39678	Amazon
Fenugreek	*Trigonella foenum-graecum*	311463	Amazon
Hazelnut	*Coylus avellana*	29539	FARRP
Hickory nut (shagbark)	*Caryra tomentosa*	33023	HickoryNuts.net
Lentil	*Lens culinaris*	337	Hy_Vee
Lima bean	*Phaseolus lunatus*	9416	FARRP
Lupin	*Lupinus albus*	46775	Amazon
Macadamia nut	*Macadamia integrifolia*	35661	FARRP
Mung bean	*Vigna radiata*	35211	Nuts.com
Pea	*Pisum sativum*	1963	Hy_Vee
Peanut (runner)	*Arachis hypogaea*	97949	FARRP
Pecan	*Carya illinoinensis*	31340	FARRP
Pine nut	*Pinus edulis*	26628	FARRP
Pinto bean	*Phaseolus vulgaris*	30670	Hy_Vee
Pistachio	*Pistacia vera*	343856	FARRP
Sesame	*Sesamum indicum*	24222	FARRP
Soybean	*Glycine max*	75126	Nuts.com
Tiger nut	*Cyperus esculentus*	32	Amazon
Walnut	*Juglans regia*	38604	Amazon
Wheat	*Triticum aestivum*	130789	FARRP

The final factor which influenced sample selection was proteome availability. A more comprehensive database is most often associated with a higher number of protein sequences for a given organism. Because amino acid identification using mass spectrometry has the highest degree of confidence when a database is complete, samples with comprehensive proteome availability such as walnut and soybean were included (Table 2.1).

Dried, raw forms of the seed were sourced from grocery stores, online vendors, and previously acquired stock.

2.4.2 Sample extraction

As outlined in **2.2 Materials and Methods**, raw, dried samples were sourced and ground into a fine powder using liquid nitrogen. This finely ground, unextracted powder will be referred to as 'ground sample' for the remainder of this document. The ground sample then underwent protein extraction in urea buffer. The resulting product will be referred to as 'extracted sample' or 'extract' for the remainder of this document. For the purposes of this experiment, 'extraction efficiency' and 'protein recovery' will be used interchangeably and are defined as the protein content as measured by 2DQ divided by total amino acids as measured by HAA for a given sample.

The buffer used for protein extraction used urea and thiourea, both of which are chaotropic and denaturing agents, to disrupt secondary protein structure and allow for solubilization of otherwise insoluble proteins. Solubilization is also increased by DTT, a reducing agent which functions to destabilize disulfide bridges. Finally, Tris was included to maintain pH stability.

Data in Table 2.2 represent the amount of soluble protein recovered by extraction compared to the total proteins in fresh weight sample. Soluble protein concentrations of the extracts varied from 0.26 μg/μL – 13.41 μg/μL and extraction efficiencies ranged from 13% - 98% with the highest occurring in peanut (Table 2.2).

The highest concentration and % recovery of protein was extracted in peanut relative to ground sample protein content (Table 2.2). A relatively large range of protein extractability was observed among seed samples with the lowest percentage of extracted proteins observed in tiger nut. Interestingly, tiger nut is not actually a seed but rather a tuber. These nodules are found along the root system of a sedge plant native to northern Africa and the Middle East, and thus, it is understandable that it would be relatively low in protein.

While a high extraction efficiency was observed in peanut, lower protein levels were recovered in other samples. Baru nut is third highest in protein in its fresh, ground state (Figure 2B); however, it is second lowest in protein recovery (Table 2.2). This indicates a sizeable abundance of proteins in Baru nut which are not soluble in the buffer used for this experiment, likely cell wall proteins. Additionally, most wheat proteins require an ethanol buffer for solubilization so a recovery of ~27% is reasonable for the proportion of proteins which are soluble in urea. Tiger nut is the lowest in total protein (Figure 2B) and presented the lowest protein recovery (Figure 2A). This should be considered when analyzing later data because a relatively small portion of the overall protein content is represented in the extracted sample.

Table 2.2: Soluble protein recovered by extraction

	Total protein in ground sample (mg/g)	% Protein Recovery	Soluble Protein Concentration (ug/uL)
Peanut	273.49	98.07	13.41
Sesame	291.41	79.49	11.58
Pistachio	204.48	77.48	7.92
Pine nut	155.08	73.33	5.69
Brazil nut	167.84	59.98	5.03
Pea	258.85	58.02	7.51
Cashew	257.70	53.75	6.93
Hickory nut	163.00	53.75	3.37
Bambara	197.63	50.21	4.96
Almond	206.63	47.70	4.93
Walnut	157.48	46.90	3.92
Hazelnut	175.70	46.39	4.08
Cow pea	252.05	43.99	5.54
Macadamia	81.18	43.94	1.78
Pinto bean	237.49	43.51	5.17
Lima bean	260.00	41.70	5.42
Chickpea	311.22	39.40	6.13
Lupin bean	418.14	34.70	7.25
Pecan	109.40	33.11	1.81
Soy	422.60	31.87	6.89
Lentil	263.30	31.23	4.11
Wheat	174.54	27.42	2.39
Fenugreek	250.58	25.53	3.20
Mung bean	299.35	23.24	3.48
Baru nut	318.08	17.53	2.79
Tiger nut	39.52	13.30	0.26

All samples listed in descending order by % Protein Recovery. '% Protein Recovery' was determined by dividing soluble protein content of the extract (mg/g FW) by fresh, ground sample protein content (mg/g FW), expressed as a percentage. 'Soluble protein concentration of the extract' was determined by 2D Quantification.

Protein extractability is a significant barrier to an experiment of this scale. We acknowledge that different protein families are optimally extractable in different buffers. Furthermore, we acknowledge that plants vary in their distribution of such protein families, thus differing in total protein extractability using the urea buffer outlined above. Mitigating this variability would require optimization of buffers for each of the 26 samples. Even with optimized buffers, 100% protein recovery would not occur, and a level of bias would still be incurred. Therefore, we acknowledge that variability is introduced by use of the same buffer for all samples, and we accept this limitation to allow for feasibility and reproducibility of the study.

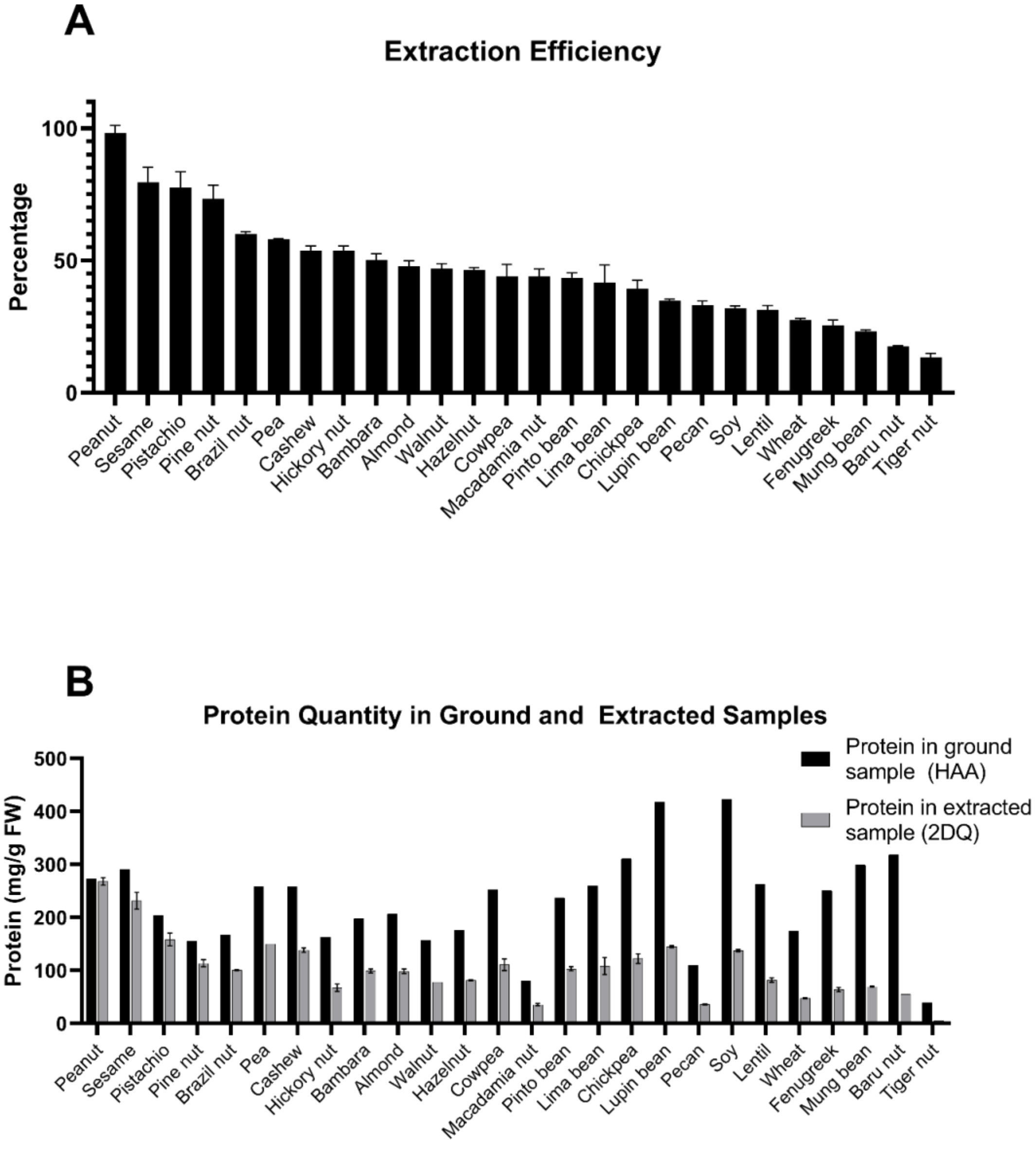


Figure 2.2 A) Extraction efficiency

Efficiency of protein extractions was determined by dividing protein (mg/g FW) as determined by HAA by that determined by 2DQ before and after extraction, respectively. 2DQ was performed on all extraction replicates (n=3); mean and SEM are presented. Total protein content in ground sample was determined for each sample by HAA.

B) Protein quantity in ground and extracted samples
Total protein quantity (mg/g FW) for ground samples using HAA and extracted samples using 2DQ. Samples are graphed in the same order as in Figure 2A for comparison.

2.4.3 Qualitative data analysis

SDS-PAGE gels were run under reducing conditions to confirm protein extraction and for qualitative assessment of protein profiles across all 26 samples (Figure 3). The extraction method was confimed by visualization of major seed storage proteins in peanut (Figure 3A).

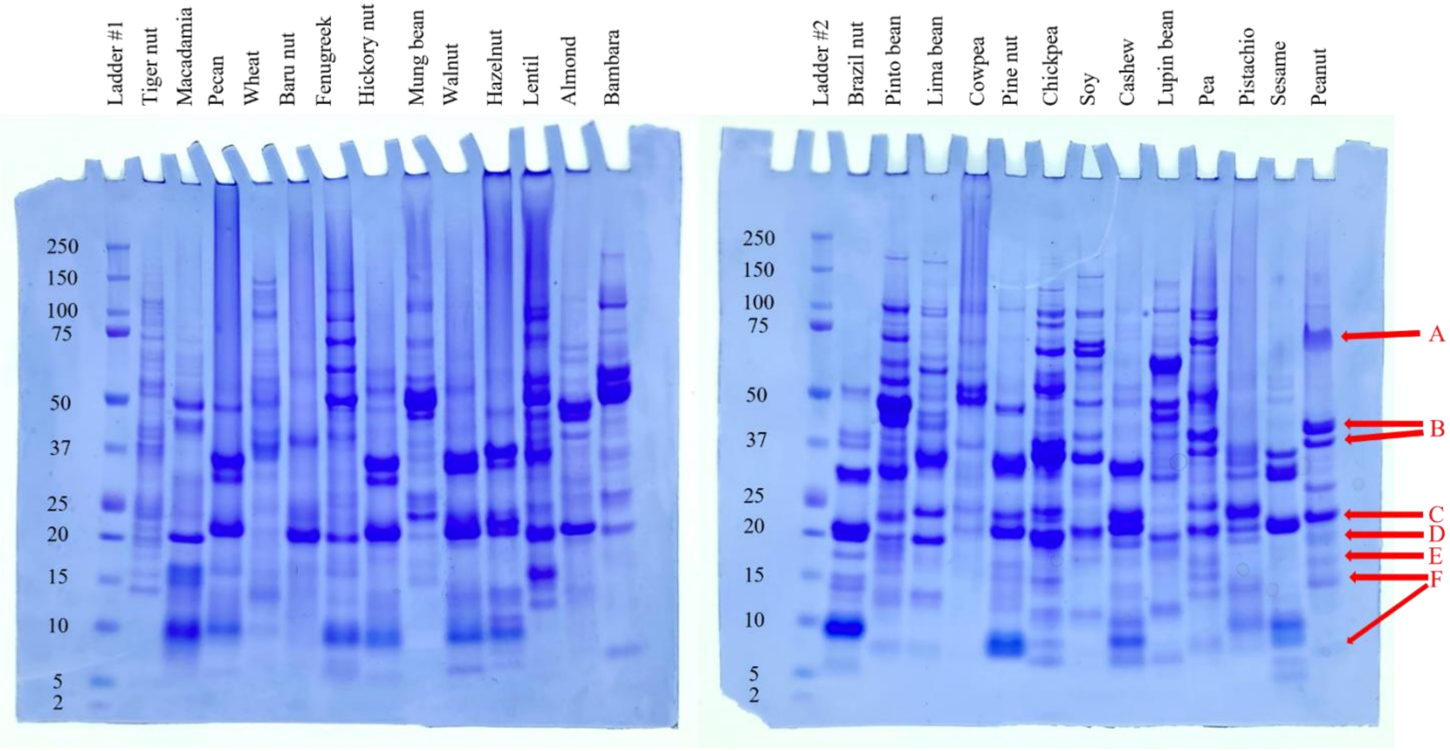


Figure 2.3: SDS-PAGE of extracted samples

Reducing SDS PAGE gels visualize the protein profiles of all 26 sample extracts. One replicate from triplicate extractions for each sample is represented. Major peanut allergens are represented including Ara Ara h 1 (A), Ara h 3 - acidic subunits (B), Ara h 3 - basic subunit (C), Ara h 2.02 (D), Ara h 2.01 (E), and Ara h 6 acidic (top) and basic (bottom) subunits (F).

2.4.4 Amino acid analysis

As briefly mentioned in **2.4.2 – Protein extraction**, HAA provided total amino acid content as well as individual amino acid abundance in each sample. The raw data can be viewed in its entirety (Ground_HAA_RawData.xls). Hydrolyzed amino acid analysis was selected rather than free amino acid analysis because it accounts for amino acids bound in proteins as well as freely available residues. For this reason, total HyP abundance as measured by HAA may account for more HyP than is present in protein molecules and peptides.

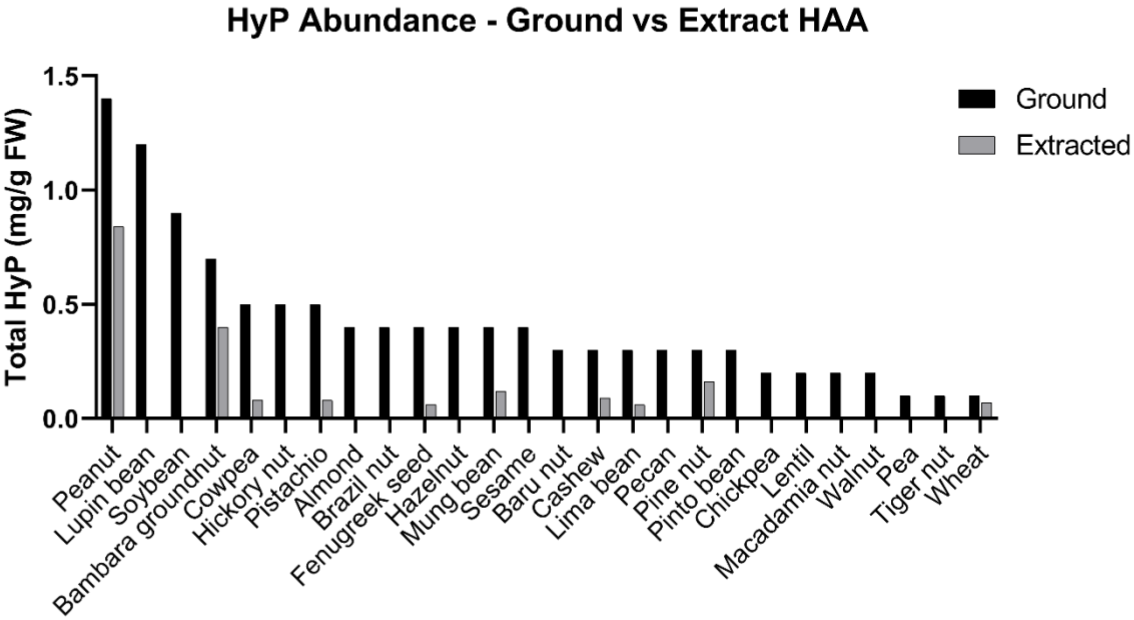


Figure 2.4: HyP abundance – ground vs extract HAA

HyP identified (mg/g FW) by HAA in ground and extracted samples. Samples which lack a bar for extracted HyP abundance tested below the limit of quantification for this analysis (6.25 pmol/µL). HyP was detected in the following sample extracts: peanut, Bambara groundnut, cowpea, pistachio, fenugreek, mung bean, cashew, lima bean, pine nut, and wheat.

HAA was performed on both ground and extracted samples. The data from HAA of ground sample detected HyP presence in each of 26 samples (Figure 2.4).

Contrastingly, HAA of extracted samples detected HyP in only 10 samples. For the other 16 samples, HyP levels were below the limit of quantification for this analysis (6.25 pmol/μL). For plotting purposes, such results will be quantified as "0"; however, we acknowledge that levels are not necessarily zero but rather are below the limit of quantification. The varying abundances of HyP in sample extracts indicate that HyP present in the ground sample exists in proteins which were not effectively extracted using a chaotropic and reducing buffer. This includes cell wall proteins as well as glutenins and gliadins in wheat among others. HyP is known to exist as part of hydroxyproline-rich glycoproteins (HRGPs) as a site of *N*-glycosylation in the cell wall[109] so it is reasonable that at least some HyP would be retained in cell wall proteins during extraction, which are insoluble using the buffer outlined in this experiment.

Because HyP is a modified proline residue, the abundance of proline residues could theoretically impact total HyP. We acknowledge that, for proline abundance to consistently impact HyP abundance, each proline would require equal susceptibility to hydroxylation. Without protein modeling, the position of prolines in the tertiary and quaternary structures cannot be verified. Proline residues on the protein's surface would have a greater susceptibility for modification while residues folded to the interior of the structure would be inhibited from interaction with the modifying enzyme. Furthermore, the activity of prolyl hydroxylases are seemingly nonspecific in soluble plant cell proteins. Without further knowledge of the factors impacting enzyme activity, it cannot be assumed that even surface proline residues are not equally susceptible to hydroxylation. Although there are limiting factors, proline and HyP abundances were assessed by HAA data both before and after extraction (Figure 2.5).

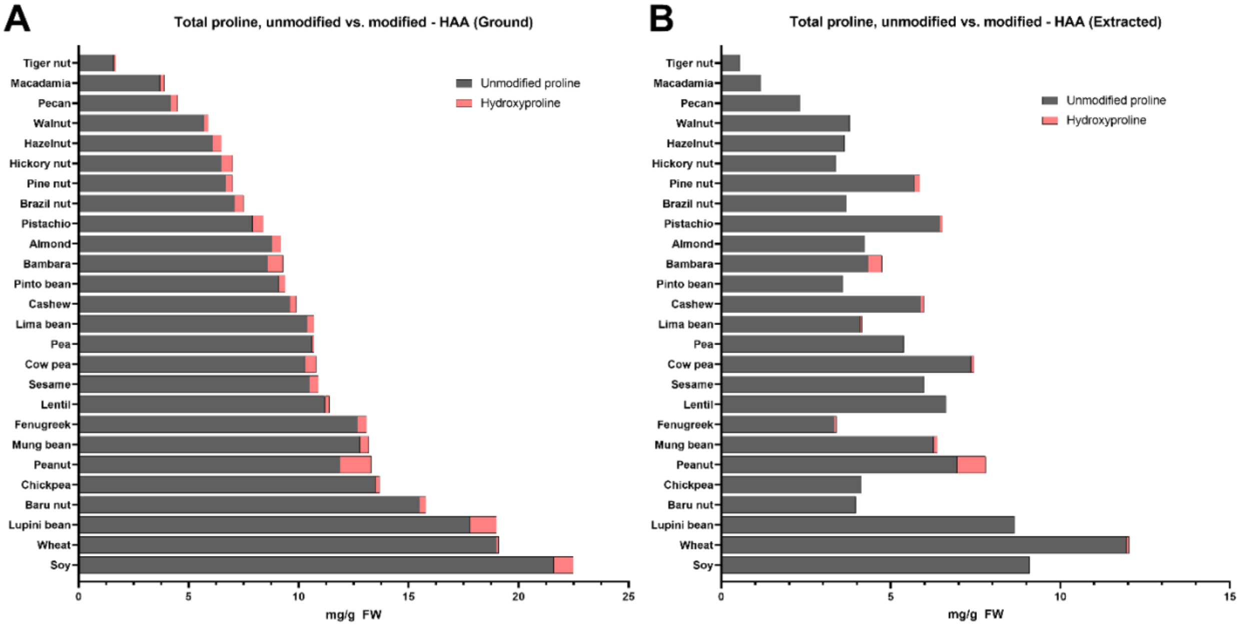


Figure 2.5: Modified and unmodified proline abundance

Unmodified proline stacked with hydroxyproline as determined by HAA of fresh, ground sample (A) and extracted sample (B). Bars which lack hydroxyproline representation in the extracted sample (B) tested below the limit of detection for HAA.

2.4.5 Correlation and outlier assessments

In ground samples, soy was found to be highest in proline and high in hydroxyproline (0.9 mg/g FW HyP) (Figure 2.5-A); however, extracted soy tested below the limit of quantification for HyP (6.25 pmol/µL) and is plotted at 0 (Figure 2.5-B) indicating that the high abundance of HyP in the ground sample largely exists in insoluble proteins. A reduction in proline and HyP occurred across all samples, as expected, due to some proteins' low extractability in the reducing and denaturing buffer. However, there was high variability in the amount of HyP and proline extracted relative to abundance before extraction. This reflects the variability in extraction efficiency

(Figure 2.2) and suggests that HyP and proline are not present at consistent proportions in soluble vs. insoluble proteins across all samples studied.

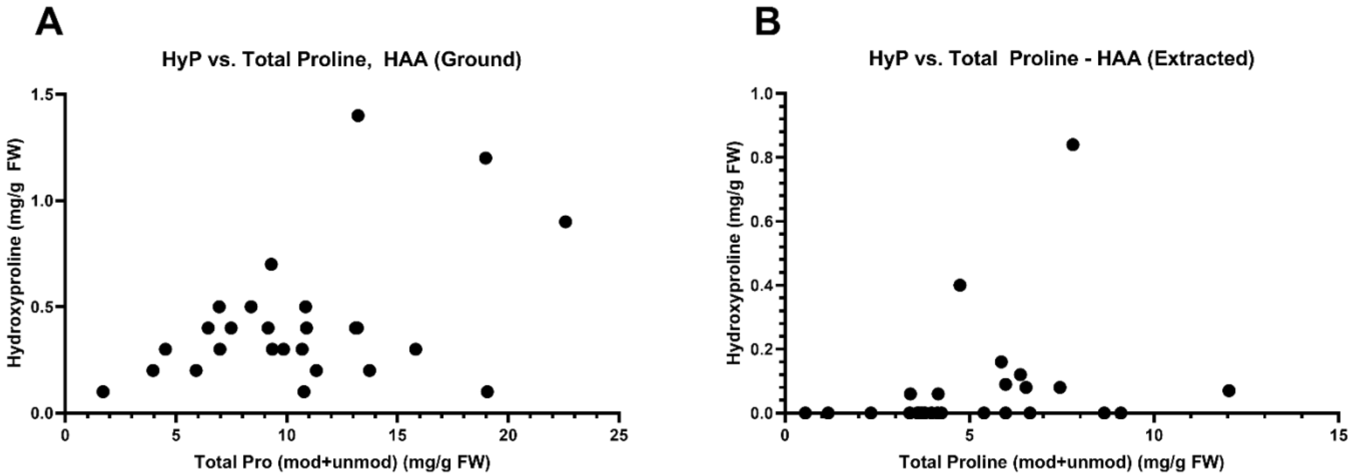


Figure 2.6: HyP vs total proline in ground and extracted samples

Hydroxyproline plotted as a function of total proline (sum of all modified & unmodified proline residue abundance) in ground (A) and extracted (B) samples as identified by HAA. Datapoints present on X-axis represent samples in which hydroxyproline tested below the limit of quantification (6.25 pmol/uL) for HAA.

Spearman's rank correlation was calculated to assess the relationship between HyP and total proline concentrations determined by HAA. In ground samples, there was a positive correlation between the two variables, r(24) = .229, p = .1301. A statistical significance for the reported positive correlation was not observed; therefore, there is insufficient evidence to reject part 1 of the null hypothesis that there is no correlation between HyP and total proline concentrations in ground samples as determined by HAA (Figure 2.6-A). In extracted samples, there was a positive correlation between the two variables, r(24) = .444, p = .0116. A statistical significance for the reported positive correlation was observed; therefore, there is sufficient evidence to reject part 2 of the null hypothesis that there is no correlation between HyP and total proline concentrations in extracted samples as determined by HAA. Though a relatively loose correlation is observed between the two variables in extracted samples, the correlation is statistically reliable (Figure 2.6-B).

Because the Spearman's ranked correlation is nonlinear, outliers have minimal impact on the correlation evaluation of the whole dataset. Therefore, correlation assessments included all outliers in this case. Still, outliers were identified to provide information about individual samples. Because the spread of the data was relatively high, a FDR of Q = 5% was selected for the ROUT method for outlier identification in both datasets. For ground samples, soy, lupin bean, and peanut were identified as outliers (Q = 5%). For extracted samples, peanut and Bambara groundnut were identified as outliers (Q = 5%). For soy and lupin, their significantly high proportion of HyP before extraction compared to their unremarkable HyP concentrations in soluble proteins indicate that much of their HyP likely exists in the cell wall.

All five outliers (peanut was counted once for each HAA) are taxonomically related as part of the legume family. Interestingly, both peanut and the Bambara groundnut are the only two samples, across all 26 studied, which share a geocarpic growing pattern. The literature is lacking in studies on HyP in underground-growing legumes and may be an interesting avenue for future research.

2.5 Conclusions

After detection of HyP in every ground sample using HAA it is evident that HyP is present in species across Viridiplantae. Levels of detection in extracted samples using the same analysis were markedly lower, indicating that much of the HyP identified in the original sample presumably exists in the cell wall (e.g. HRGPs) or in other proteins which are not extractable in a chaotropic and reducing buffer. HAA also indicated total proline levels which did not correlate with HyP levels in ground samples but did correlate in extracted samples. The lack of correlation between HyP and proline presence across all ground samples evaluated (Figure 2.6-A) may indicate inconsistency in total HyP abundance across Viridiplantae. Conversely, the moderate correlation between soluble HyP and total proline indicates that there is some relationship between soluble HyP and soluble proline.

Interestingly, two samples (peanut and Bambara groundnut) proved to be substantially higher in hydroxyproline than others, even when normalized to total proline content. Discuss further/speculate what might contribute to high soluble HyP abundance.

HyP abundance in each sample, while helpful, is not sufficient to explore patterns of hydroxylation. To further investigate hydroxyproline presence using more specific methods, each sample extract was then studied using mass spectrometry as discussed in detail in the following chapter.

REFERENCES

(1) “Allergies and Hay Fever,” Centers for Disease Control and Prevention, 2018.

(2) Gupta, R. S.; Warren, C. M.; Smith, B. M.; Jiang, J.; Blumenstock, J. A.; Davis, M. M.; Schleimer, R. P.; Nadeau, K. C. Prevalence and Severity of Food Allergies Among US Adults. *JAMA network open* **2019,** *2* (1), e185630.

(3) Peters, R. L.; Koplin, J. J.; Gurrin, L. C.; Dharmage, S. C.; Wake, M.; Ponsonby, A.-L.; Tang, M. L. K.; Lowe, A. J.; Matheson, M.; Dwyer, T.et al. The prevalence of food allergy and other allergic diseases in early childhood in a population-based study: HealthNuts age 4-year follow-up. *Journal of Allergy and Clinical Immunology* **2017,** *140* (1), 145.

(4) Skolnick, H. S.; Conover-Walker, M. K.; Koerner, C. B.; Sampson, H. A.; Burks, W.; Wood, R. A. The natural history of peanut allergy. *Journal of Allergy and Clinical Immunology* **2001,** *107* (2), 367.

(5) Koppelman, S. J.; Wensing, M.; Ertmann, M.; Knulst, A. C.; Knol, E. F. Relevance of Ara h1, Ara h2 and Ara h3 in peanut-allergic patients, as determined by immunoglobulin E Western blotting, basophil–histamine release and intracutaneous testing: Ara h2 is the most important peanut allergen. **2004,** *34* (4), 583.

(6) Bernard, H.; Guillon, B.; Drumare, M. F.; Paty, E.; Dreskin, S. C.; Wal, J. M.; Adel-Patient, K.; Hazebrouck, S. Allergenicity of peanut component Ara h 2: Contribution of conformational versus linear hydroxyproline-containing epitopes. *The Journal of allergy and clinical immunology* **2015,** *135* (5), 1267.

(7) Palmer, G. W.; Dibbern, D. A., Jr.; Burks, A. W.; Bannon, G. A.; Bock, S. A.; Porterfield, H. S.; McDermott, R. A.; Dreskin, S. C. Comparative potency of Ara h 1 and Ara h 2 in immunochemical and functional assays of allergenicity. *Clinical immunology (Orlando, Fla.)* **2005,** *115* (3), 302.

(8) Üzülmez, Ö.; Kalic, T.; Mayr, V.; Lengger, N.; Tscheppe, A.; Radauer, C.; Hafner, C.; Hemmer, W.; Breiteneder, H. The Major Peanut Allergen Ara h 2 Produced in Nicotiana benthamiana Contains Hydroxyprolines and Is a Viable Alternative to the E. Coli Product in Allergy Diagnosis. **2021,** *12*.

(9) Chatel, J. M.; Bernard, H.; Orson, F. M. Isolation and characterization of two complete Ara h 2 isoforms cDNA. *International archives of allergy and immunology* **2003,** *131* (1), 14.

(10) De Leon, M. P.; Glaspole, I. N.; Drew, A. C.; Rolland, J. M.; O'Hehir, R. E.; Suphioglu, C. Immunological analysis of allergenic cross-reactivity between peanut and tree nuts. *Clinical & Experimental Allergy* **2003,** *33* (9), 1273.

(11) Migueres, M.; Dávila, I.; Frati, F.; Azpeitia, A.; Jeanpetit, Y.; Lhéritier-Barrand, M.; Incorvaia, C.; Ciprandi, G. Types of sensitization to aeroallergens: definitions, prevalences and impact on the diagnosis and treatment of allergic respiratory disease. *Clinical and translational allergy* **2014,** *4*, 16.

(12) Du Toit, G.; Katz, Y.; Sasieni, P.; Mesher, D.; Maleki, S. J.; Fisher, H. R.; Fox, A. T.; Turcanu, V.; Amir, T.; Zadik-Mnuhin, G.et al. Early consumption of peanuts in infancy is associated with a low prevalence of peanut allergy. *Journal of Allergy and Clinical Immunology* **2008,** *122* (5), 984.

(13) Lack, G. Epidemiologic risks for food allergy. *Journal of Allergy and Clinical Immunology* **2008,** *121* (6), 1331.

(14) Goodman RE, E. M., Ferreira F, Sampson HA, van Ree R, Vieths S, Baumert JL, Bohle B, Lalithambika S, Wise J, Taylor SL; Mol. Nutr. Food Res, 2016.

(15) In *Public Law 117-11*; Congress, t., Ed., 2021.

(16) In *21 USC 301 note*, 2004.

(17) Bublin, M.; Breiteneder, H. Cross-reactivity of peanut allergens. *Curr Allergy Asthma Rep* **2014,** *14* (4), 426.

(18) Mennini, M.; Dahdah, L.; Mazzina, O.; Fiocchi, A. Lupin and Other Potentially Cross-Reactive Allergens in Peanut Allergy. *Current Allergy and Asthma Reports* **2016,** *16* (12), 84.

(19) Ajilogba, C. F.; Olanrewaju, O. S.; Babalola, O. O. Improving Bambara Groundnut Production: Insight Into the Role of Omics and Beneficial Bacteria. **2022,** *13*.

(20) Tan, X. L.; Azam-Ali, S.; Goh, E. V.; Mustafa, M.; Chai, H. H.; Ho, W. K.; Mayes, S.; Mabhaudhi, T.; Azam-Ali, S.; Massawe, F. Bambara Groundnut: An Underutilized Leguminous Crop for Global Food Security and Nutrition. *Frontiers in nutrition* **2020,** *7*, 601496.

(21) Shimizu, M.; Igasaki, T.; Yamada, M.; Yuasa, K.; Hasegawa, J.; Kato, T.; Tsukagoshi, H.; Nakamura, K.; Fukuda, H.; Matsuoka, K. Experimental determination of proline hydroxylation and hydroxyproline arabinogalactosylation motifs in secretory proteins. **2005,** *42* (6), 877.

CHAPTER 3: IDENTIFICATION AND TIERED EVALUAITON OF HYDROXYPROLINE SITES DETECTED USING DATA-DEPENDENT ACQUISITION

3.1 Introduction

3.1.1 Hydroxyproline (HyP) in plant proteins

Posttranslational modifications (PTMs) have been known to have substantial impact on IgE binding epitopes. Specifically, PTMs on linear epitopes can be studied by analyzing a protein's primary structure. An epitope consisting of 10 amino acid residues in timothy grass pollen allergen, Phl p 1, is known to contain two hydroxyproline (HyP)[72,110]. These two sites, found in naturally-occurring Phl p 1, impact levels of IgE reactivity[72].

The sera of over 90% of peanut-allergic individuals present IgE binding to peanut allergen, Ara h 2[65]. Both isoforms, Ara h 2.01 and Ara h 2.02, present a linear, HyP-containing motif which has been shown to influence IgE binding as well as mediator release in humanized rat basophil leukemia (RBL) cell assays[63,67-69].

To explore HyP presence in soluble plant cell proteins, 26 species from across Viridiplantae were selected for proteomic analysis. In the previous chapter, identification of HyP and its abundance was evaluated by hydrolyzed amino acid analysis (HAA). While it is helpful to understand the abundance of HyP compared to other amino acids in the sample, data from HAA provides few details compared to other methods.

3.1.2 Mass spectrometry (MS)

Mass spectrometry (MS) involves ionization and the detection of mass-to-charge ratios (m/z) to deduce molecular makeup. A common use for MS is to characterize protein molecules. MS analysis of proteins allows for the study of primary amino acid structure. This information is invaluable when examining potentially or known allergenic proteins because linear motifs are defined by their amino acid sequences.

The primary structure of protein molecules can be examined when data-dependent acquisition (DDA) mode is enabled, a process for which protein digestion is a prerequisite. Chromatographic separation techniques are often coupled to MS to reduce the biological complexity of a sample, optimize ionization and detection, and increase reproducibility of the experiment[111].

Each peptide detected by the mass analyzer will be recorded according to its retention time as well its m/z. During each cycle of acquisition, the ionized peptides first detected are referred to as the “precursor ions”. The spectrum obtained from each precursor ion will be referred to as the “MS1” spectrum. From all precursor ions in a cycle, the most abundant ions are selected for fragmentation. When a precursor ion is fragmented, multiple “fragment ions” are produced. For each fragment ion, a “MS2” spectrum is generated, essentially providing a more detailed look at that ion’s respective MS1.

3.1.3 MS software analysis

Raw MS files are then uploaded to MS software for processing. The software functions to overlay the m/z of fragment ions to mathematically deduce the potential monoisotopic mass of amino acid positions, thus allowing for residue identification. More complete fragmentation of precursor ions allows for greater confidence in amino acid sequence characterization, while sparse fragmentation leaves a greater number of possible combinations, thus reducing sequence confidence. Furthermore, PEAKS software for MS data analysis allows for characterization of primary amino acid structure by comparing experimental MS data with that derived from a protein database. This provides a pre-determined list of m/z values with which PEAKS can match experimental m/z for increased sequence confidence.

3.1.4 MS for posttranslational modification (PTM) detection

Just as amino acids are identified by the overlap of fragment ions for a precursor ion, the software can also identify posttranslational modifications (PTMs) by scanning for the known mass of a residue plus the known residue of the modification of interest. Successful identification occurs when a fragment overlap matches the mass of a residue plus its modification.

Many times, PTMs are reversible and function to turn protein activity on or off[112]. Proline hydroxylation, unlike many PTMs, is the permanent addition of a hydroxyl group to the 3˚ or 4˚ carbon resulting in a 16 atomic mass unit increase to the molecule[112].

Because PTMs are often lacking in proteomic databases and therefore may not appear on traditional database searches, the PTM Profile in PEAKS integrates database search results with *de novo* sequencing results to acquire the most accurate prediction of PTM positions and relative abundance[113].

3.1.5 MS manual analysis

One significant limitation to site determination is that there are only two options available for determination of confident PTM sites: minimal ion intensity and Ascore. While these are helpful, they are multiple variables integrated into single thresholds for confidence. This disallows customization of confidence parameters. Even though these confidence parameters are strong, well-rounded evaluations for PTM identification, they are limited by their lack of flexibility. With only minimum ion intensity and Ascore algorithms for confidence evaluation, we were unable to customize an evaluation method by increasing weight of other criteria.

We sought to further refine the number of sites to only the most confident as well as adjusting the weights placed on each variable, namely minimum ion intensity and Ascore as determined by the PEAKS algorithm. Instead, emphasis was placed on a refined residue window. We determined to enforce a qualification of MS2 spectra for fragments which overlap directly on either side of the HyP residue with one maximum adjacent residue. We felt this provided the best balance between strict confidence criteria while being flexible enough to provide a reasonable number of datapoints. Because this type of analysis customization is not available in PEAKS, it was performed manually using Microsoft Excel (MS Excel) macros with Visual Basic for Applications (VBA). Furthermore, when only *de novo* sequencing is available, ALC is the only confidence measurement readily available to assess peptides containing PTMs. While it is better than no measurement at all, ALC is certainly too vague for use in drawing any meaningful conclusions. For this reason, MS Excel macros using VBA were designed to identify the local score of each residue position in peptide MS1.

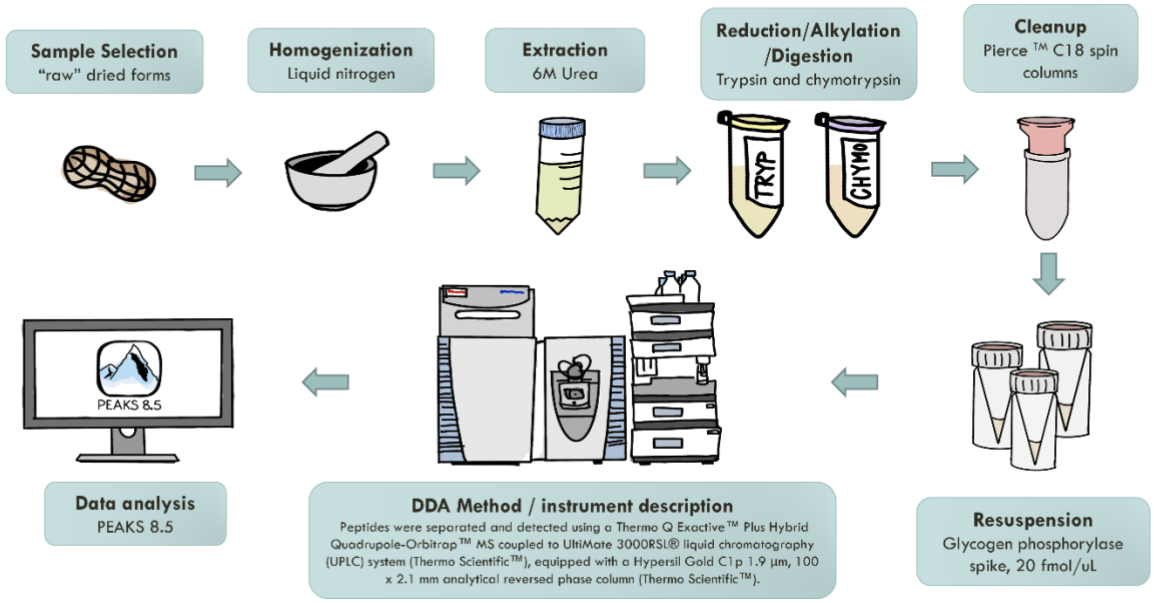
Within Tier 1, an advanced sub-degree of confidence was given to sites found in both tryptic and chymotryptic experiments (Table 3.3). We acknowledge that the sites found using chymotryptic and tryptic enzymes require the HyP site to occur at a position close enough to a respective cut site (or really two cut sites, one on the N- terminal side of HyP and one on C- terminal side of HyP) that produce a peptide that 1.) has adequate ionization efficiency for strong flight in the instrument and 2.) is small enough to allow for near complete or complete MS/MS fragmentation. HyP residues of interest may or may not occur in such positions and the degree to which this occurs is undetermined. Thus, variability is introduced across samples, a concession deemed necessary to perform a universal study. Furthermore, by including sites identified in chymotryptic samples, this variability is mitigated because an increased number of cut sites results in greater number of peptides for comparison. This allows for greater coverage of a sample's protein profile during proteomic analysis.

3.2 Materials and Methods

3.2.1 Sample Preparation

3.2.1.1 Homogenization and extraction

Methods for sample homogenization using liquid nitrogen and triplicate extractions extracted in triplicate in a reducing and denaturing buffer of 6M urea (BioRad), 2M thiourea (Sigma-Aldrich), 20 mM dithiothreitol (DTT) (Sigma-Aldrich), 50 mM Tris (Sigma-Aldrich) pH 8.8 buffer at 50 mg/mL (w/v) are explained in detail in (**2.3 Materials and Methods**). An illustrated workflow of the sample preparation, acquisition, and software analysis is displayed in Figure 3.1.



Peptides were separated and detected using a Thermo Q Exactive™ Plus Hybrid Quadrupole-Orbitrap™ MS coupled to UltiMate 3000RSL® liquid chromatography (UPLC) system (Thermo Scientific™), equipped with a Hypersil Gold C1p 1.9 µm, 100 x 2.1 mm analytical reversed phase column (Thermo Scientific™).

Figure 3.1: Methods workflow

The process of sample preparation, data acquisition, and initial analysis is summarized.

3.2.1.2 Digestion:

All digestion protocols used ultrapure LC-MS water.

For tryptic digests, the reducing buffer contained 100 mmol DTT, alkylation buffer contained 50 mmol IAA, and digestion buffer contained 50 mM ammonium bicarbonate. For chymotryptic digests, 500mM DTT was used for reduction, 500 mM IAA was used for alkylation, and 500 mM Tris-HCl (pH 8.0), 10mM CaCl was used for digestion. Protein content determination (2D Quantification, Cytiva) ensured the sample protein concentration was within the enzyme's digestion capacity: trypsin (1µg trypsin/25µg protein) and chymotrypsin (1µg chymotrypsin/20µg protein). All triplicate extractions were digested with both trypsin and chymotrypsin. Chymotrypsin was resuspended to 1 µg/µL and was stored at -20C in 10µL aliquots.

3.2.1.3 Cleanup:

All cleanup protocols used ultrapure LC-MS water.

Pierce C-18 spin columns were used to desalt digests. The official protocol was followed for sample preparation (3:1 sample to sample-buffer ratio), column activation, column equilibration, binding, wash (three washes rather than two due to the use of 6M urea extraction buffer), and elution.

3.2.1.4 Resuspension:

All resuspension protocols used ultrapure LC-MS water.

Lyophilized muscle form, rabbit glycogen phosphorylase was resuspended to 200 fmol/uL stock glycogen phosphorylase (GlyP) and stored at -20˚C. GlyP was spiked into each resuspension to act as an internal standard protein; this could be used to normalize ion intensities and assist with label free quantification (LFQ) if desired in later analyses.

GlyP stock (200 fmol/µL) was added to resuspension buffer (5% ACN, 0.1% FA) at a concentration of 20 fmol/µL. A loading volume of 2µL on column resulted in 40 fmol GlyP on column for each injection.

3.2.2 Data Acquisition

During data acquisition, samples were injected on column in triplicate to assess and ensure reproducibility of the experiment. This resulted in nine replicates for each species when digested with trypsin and nine replicates when digested with chymotrypsin. Microflow liquid chromatography was performed for separation of tryptic and chymotryptic peptides using UltiMate 3000RSL® liquid chromatography (UPLC) system (Thermo Fisher Scientific) with a Hypersil Gold C18 1.9 µm, 100 x 1.1 mm analytical reversed phase column (Thermo Fisher Scientific). Two mobile phases were used during elution. Solvent A contained 99.9% (v/v) water and 0.01% formic acid (FA). Solvent B contained 99.9% (v/v) acetonitrile (ACN) and 0.01% FA. All samples were injected at a volume of 2 µL on column where a mobile phase flow rate of 0.060 mL/min was used at Solvent B concentrations increasing from 2% – 98% over a 76-minute gradient.

Data were collected using a Thermo Q Exactive™ Plus Hybrid Quadrupole-Orbitrap™ MS (Thermo Fisher Scientific) set to data-dependent mode. MS scans of precursor ions were acquired at a resolution of 70,000 from 400 – 1400 m/z with an automatic gain control (AGC) target of 3 x 10^6 and a maximum injection time (IT) of 100 ms. Up to the top 20 most abundant precursor ions with charges of 2, 3, or 4 in each MS scan were selected for fragmentation using higher-energy collisional dissociation (HCD) resulting in MS/MS spectra acquired at a resolution of 70,000 from 200 – 2000 m/z with an AGC target of 1 x 10^5, maximum IT of 240 ms, isolation window of 2.0 m/z and isolation offset of -0.4 m/z. To reduce repeated fragmentation and increase efficiency of peptide identification, a dynamic exclusion window of 20 s was enforced.

3.2.3 Software data analysis

3.2.3.1 Database selection/compilation

Where available, databases were compiled from UniProt to maximize consistency. When available, the reference proteome was downloaded; if no reference proteome was available, the proteome with the highest protein count was downloaded.

Proteomes with more than 20,000 sequences for the native species will be referred to as "substantial databases" for the remainder of the study. All proteins were downloaded in the FASTA (isoform + canonical) form.

For samples without reference proteomes or substantial databases, sequences from SwissProt, UniProt, NCBI, and our own lab information were combined to increase the robustness of protein databases used in this study. Protein sequences for lima bean (*Phaseolus lunaus*) and pistachio (*Pistachia vera*) were acquired by running AUGUSTUS gene prediction on the samples' respective reference genomes using *Arabidopsis thaliana* as the model organism, genes reported on both strands, no alternative transcripts, and structure allowing for the prediction of any number of genes, including partials[114,115]. Sequences for cashew were also originally acquired using AUGUSTUS[116]. For species with mRNA transcription sequences available through the Hardwood Genomics Project[117], databases were created by selecting the species of interest and creating a downloadable collection of mRNA-polypeptide with the following information reported: time last modified, name, identifier, protein sequence, relationship, organism, and feature publication.

3.2.3.2 PEAKS 8.5 – Settings

Raw LC-MS/MS files (Xcalibur®, ThermoFisher Scientific) were uploaded directly to PEAKS 8.5 where *de novo* analyses were performed on each group of nine replicates. The following settings were applied: 5 ppm and 0.06 Da error tolerance, max 3 variable PTM per peptide allowed, and up to 5 candidates reported per spectrum. Following *de novo* analyses, PEAKS database (DB) searches were performed on each group using curated databases comprised of available reference and predicted protein sequences. The following settings for PEAKS DB searches were applied: 5 ppm and 0.02 error tolerance, non-specific cleavage allowed at one end of the peptide, maximum of 1 missed cleavage allowed per peptide, and a maximum of 3 allowed variable PTM per peptide. The common Repository of Adventitious Proteins (cRAP) was used as the contaminant database.

For samples which lacked a reference or substantial proteome, reference genome (for genome-derived transcriptome prediction), and a closely related organism that possesses a reference or substantial proteome, and which had ≤ 2,000 native sequences on UniProt, *de novo* sequencing alone was performed to predict peptide and protein primary structure. Samples which were analyzed using only *de novo* sequencing include Brazil nut (*Bertholletia excelsa*), lentil (*Lens culinaris*), and tiger nut (*Cyperus esculantus*).

The dataset was first filtered by HyP modifications, including only peptides which contain a HyP modification identified by the PEAKS 8.5 software. As stated previously, this dataset only included peptides identified at a false discovery rate (FDR) of ≤ 1%. First, all peptides were filtered by FDR at 1%, A Score (for PTMs) ≥ 20, protein at -10lgP = 20, and *de novo*-only peptides ≥ 80% ALC.

The data quality is limited by the quality of the database against which the spectra are compared. For this reason, a tiered evaluation system was developed to accurately assess sites identified by databases of different qualities according to our criteria. It is important to note that the peptides and samples are not the subject of evaluation; rather, the HyP residue sites themselves are evaluated. This allows sites within an organism to be identified by sequences from different source databases and each would receive a score based on the site's respective associated database.

3.2.3.3 Evaluation Tiers Defined (Table 3.1):

Tier 1: The top tier (Tier 1) comprises the highest-quality HyP sites according to our parameters. Tier 1 sites have been identified by cross-reference to a protein sequence from a reference or substantial proteome native to the species of study. Because of the native sequence reference, we can be highly confident that the HyP site 1.) exists in that specified peptide and 2.) occurs at that position within the peptide. Samples for which a reference proteome was available include almond (*Prunus dulcis*), chickpea (*Cicer arietinum*), lupin bean (*Lupinus albus*), mung bean (*Vigna radiata*), peanut (*Arachis hypogaea*), pinto bean (*Phaseolus vulgaris*), sesame (*Sesamum indicum*), soybean (*Glycine max*), walnut (*Juglans regia*), and wheat (*Triticum aestivum*). For one sample, cowpea (*Vigna unguiculata*), UniProt listed 39,000 protein sequences but lacked a reference proteome. Because the volume of native proteins was large, it will be referred to as a "substantial" protein database for that sample. HyP sites identified using sequences from a substantial protein database will be classified as Tier 1 sites.

Tier 2: The second tier (Tier 2) contains the second-highest quality HyP sites according to our parameters. Tier 2 sites have been identified by cross-reference to a predicted protein sequence from a genome-derived predicted proteome native to the species of the study. Because it is a native sequence but is predicted rather than a reference sequence it is a tier below those identified by reference proteome sequences.

Tier 3: The third tier (Tier 3) is assigned to HyP sites of slightly lower quality as compared to those of the first and second tiers (Tier 1 and Tier 2, respectively) according to our parameters. Tier 3 sites have been identified by a cross reference to a protein sequence from a reference proteome or an organism which shares a genus with that of the studied species. Because the sequence is only related and not native to the studied species, there is lower degree of confidence associated with the identified HyP site.

Tier 4: The fourth tier (Tier 4) is comprised of HyP sites with the lowest confidence of those identified in this study. These sites were identified using *de novo* sequencing. *De novo* sequencing produces potential peptide sequence characterization by *in-silico* overlaying of precursor and fragment ions detected during acquisition. The sites (and peptides to which they belong) are not associated with a specific protein but are instead independently identified peptides. These data provide insight of potential primary structure and location of HyP sites but do so at a lower level of confidence than sites in other tiers.

Tier 1	Tier 2	Tier 3	Tier 4
Sites ID'd via cross-reference with native reference or substantial proteome	Sites ID'd via cross-reference with genome-derived predicted proteome	Sites ID'd by cross-reference with related reference proteome (nonnative)	Sites ID'd via *de novo* sequencing
MS1: mass error 5ppm MS2: mass error 0.02 Da Peptide FDR: 1%, Protein -10lgP ≥ 20, ALC ≥ 80%	MS1: mass error 5ppm MS2: mass error 0.02 Da Peptide FDR: 1%, Protein -10lgP ≥ 20, ALC ≥ 80%	MS1: mass error 5ppm MS2: mass error 0.02 Da Peptide FDR: 1%, Protein -10lgP ≥ 20, ALC ≥ 80%	MS1 mass error: 5ppm MS2 mass error: 0.06 Da ALC ≥ 80%

Table 3.1: Definition of tiers

Tiers are listed in descending order and defined by the quality of the sequence in a protein database which identifies a putative HyP site.

3.2.3.4 Site identification and compilation

A list of HyP sites for each sample was compiled based on PEAKS DB (settings listed in 3.2.3 – Data Analysis: PEAKS 8.5 Settings). Each putative site fell into one of the four tiers of confidence which was determined by the database which identified the site. The threshold for sites included in each of the first three tiers was an AScores equal to or exceeding 20. All samples which were searched against a database (i.e. samples with sites that fall into Tiers 1-3) were analyzed using the “peptide.csv” export file. Total detectable proline residues were also recorded for each sample using the same export file.

For analysis of *de novo*-sequenced samples (i.e. Tier 4), the “de novo peptide.csv” export was used. This export was filtered to contain only peptides which possessed a ≥ 80% ALC score. Again, quantities of total proline residues were recorded for each sample.

3.2.4 Manual data analysis

Manual analyses used the PSMions.csv export file from PEAKS. MS Excel macros were designed to determine fragmentation patterns respective to precursor ions. Fragment ions for each precursor were examined as a group to assess for overlapping ions which resulted in fragmentation on both sides of the HyP site. To allow for a reasonable number of datapoints, the criteria also allowed one adjacent residue to the HyP in the overlapping window. For the remainder of this document, “bilateral fragmentation” will refer to the overlap of fragment ions in such a way that there is fragmentation on the *N*- and *C*- terminal sides of the HyP residue with a maximum allowance of one adjacent residue within the fragment overlap.

This method of analysis yielded a list of peptides containing HyP sites of varying confidence. The quality of each site was assessed according to the database which identified that respective site. If a HyP residue was identified on a different peptide (i.e. any other residues on the peptide were unique or made the peptide unique), the HyP site was counted as a unique site. For instance, if the site occurred on a peptide in which one subsequent proline was hydroxylated and one was not, the same residue position that contained HyP in both instances would be counted for each unique peptide on which it was found.

3.2.5 Statistical analysis

To assess correlation between multiple variables across MS data exported through PEAKS, Spearman's ranked correlation and Pearson's correlation for linearity were performed using GraphPad Prism 9.1.0 for Windows, GraphPad Software, San Diego, California, USA, www.graphpad.com. Additionally, ROUT outlier tests were performed using GraphPad Prism 9.1.0 for Windows, GraphPad Software, San Diego, California, USA, www.graphpad.com.

3.3 Results and Discussion

3.3.1 HyP identification by mass spectrometry software

Each sample was prepared and injected on column for liquid chromatography coupled to tandem mass spectrometry (LC-MS/MS). These data were acquired in DDA mode which selects the top 20 most abundant ions per scan for MS/MS fragmentation.

Table 3.2: Sequences and residue identification quantities per sample

#	Sample	**Database sequences**	**Total peptides identified**	**HyP residues**	**Proline residues**
1	Almond	32104	1380	34	1193
2	Chickpea	30798	2171	36	2000
3	Cowpea	39678	2296	39	1899
4	Lupin	46775	2264	26	1845
5	Mung bean	35211	1654	34	1335
6	Peanut	97949	1513	122	1447
7	Pinto	30670	1230	12	1107
8	Sesame	24222	158	1	117
9	Soy	75126	2341	39	2177
10	Walnut	38604	479	12	420
11	Wheat	130789	1648	24	1632
12	Cashew	130789	1119	20	891
13	Lima bean	9416	212	5	201
14	Hazelnut	29539	438	0	370
15	Hickory	33023	924	36	754
16	Pistachio	343856	1055	31	899
17	Macadamia nut	35661	627	6	490
18	Pecan	31340	714	11	625
19	Brazil nut	2497	119	3	110
20	Pea	134128	2644	41	2150
21	Pine nut	26628	493	12	406
22	Fenugreek	311463	1202	44	853
23	Bambara groundnut	108636	1276	26	978

All samples from Tiers 1-3 are listed because these tiers comprise samples which were run against a database. Samples 1-11 belong to Tier 1, 12-20 to Tier 2, and 21-23 to Tier 3.The cumulative number of sequences compiled for each respective sample is listed under “DB Sequences”. “Total peptide IDs”, “HyP IDs”, and “Proline residues” were all identified using “peptide” export file from PEAKS 8.5. Note: curated databases for these samples contain sequences from multiple sources.

This provides a high volume of information for the most abundant ions while forgoing data collection on less abundant ions. Though there are advantages and disadvantages to this method, it is the most appropriate mode to use in the context of food allergens because proteins of allergenic interest typically have relatively high abundance.

Of the 26 samples in the study, there were 22 for which some type of protein database was available including a native reference proteome, genome-derived proteome predictions, or a substantial database for a related organism. After acquisition, *in silico* analysis was performed according to settings previously described (See **2.2.3 – Data Analysis, <u>PEAKS 8.5 – Settings</u>**). The sizes of respective protein databases are listed in Table 3.3 along with total number of identified peptides, HyP residues, and proline residues from trypsin-digested samples. Data acquired from chymotrypsin-digested samples were used as secondary confirmation of identified sites; however, tryptic samples were used for all initial data analysis.

We acknowledge that the quality of the database greatly impacts the degree of confidence which can be held in each protein, peptide, or residue identification. For this reason, each identified site was subjected to a tier-based evaluation system (See **Figure 1** in **3.2.3 – Data Analysis**). Number of sites for each threshold of confidence are as follows: Tier 1 – 379 sites; Tier 2 – 153; Tier 3 – 82; Tier 4 – 2,828. The full breakdown of sites per sample can be seen in Table 3.3. Because bespoke databases were curated for each sample, some species' sequence databases allowed for site identification in multiple tiers. For the entirety of the study, sites themselves are evaluated. That is to say that the sample is not defined to a tier; rather, each site belongs to a tier based on the sequence used to identify that respective site. While data are most comparable with their respective tiers, Tiers 1-3 all involve PEAKS DB searches and are relatively comparable. Tier 4 sites, those identified by *de novo*, are meant to be used as baseline information and are not comparable with site numbers from PEAKS DB searches. For this reason, some comparisons will be made within a tier and others will be made across all sites in Tiers 1-3.

Table 3.3: Site quantities by tier

Tier 1 – Native reference proteome			
Sample	**Total peptides**	**Total HyP residues**	**Chymotrypsin-confirmed sites**
Almond	1380	34	1
Chickpea	2171	36	-
Cowpea	2296	39	-
Lupin	2264	26	1
Mung bean	1654	34	-
Peanut*	1513	122	8
Pinto	1230	12	-
Sesame	158	1	-
Soy*	2341	39	-
Walnut	479	12	-
Wheat	1648	24	-
Tier 2 – Genome-predicted protein sequences			
Sample	**Total peptides**	**Total HyP residues**	
Cashew	1119	20	
Lima bean	212	5	
Hazelnut	438	0	
Hickory	924	36	
Pistachio	1055	31	
Macadamia nut	627	6	
Pecan	714	11	
Brazil nut *	119	3	
Pea	2644	41	
Tier 3 – Related species' proteomes			
Sample	**Total peptides**	**Total HyP residues**	
Pine nut	493	12	
Fenugreek*	1202	44	
Bambara groundnut*	1276	26	
Baru	54	0	
Tier 4 – *De novo* sequencing			
Sample	**Total peptides**	**Total HyP residues (≥80%ALC)**	
Bambara	6497	281	
Brazil	3756	302	
Fenugreek	5907	286	
Tiger nut	1747	116	
Peanut	5614	342	
Soy	6209	329	
Lentil	8125	517	
Pea	11068	374	
Baru	3228	281	

Total number of peptides and HyP residues identified by 'peptide' (Tiers 1-3) or 'de novo peptides' (Tier 4) exports using MS software analysis. These data represent number of putative HyP sites rather than intensity or abundance of sites.

* Samples were analyzed using *de novo* sequencing in addition to a database search

3.3.2 Investigation of factors affecting the number of identifiable HyP sites

Data from HAA of extracted samples were used for comparison against proteomic analyses because proteomics were performed on extracted samples. HyP abundances as determined by HAA (extracts) were plotted against total HyP residues identified by mass spectrometry. It is noteworthy that, because DDA only selects the top 20 ions per cycle, it is likely that HyP residues present on lesser abundant peptides would not be detected. This is not to say that they are not present, only that they would not be identified by the methods used in this experiment.

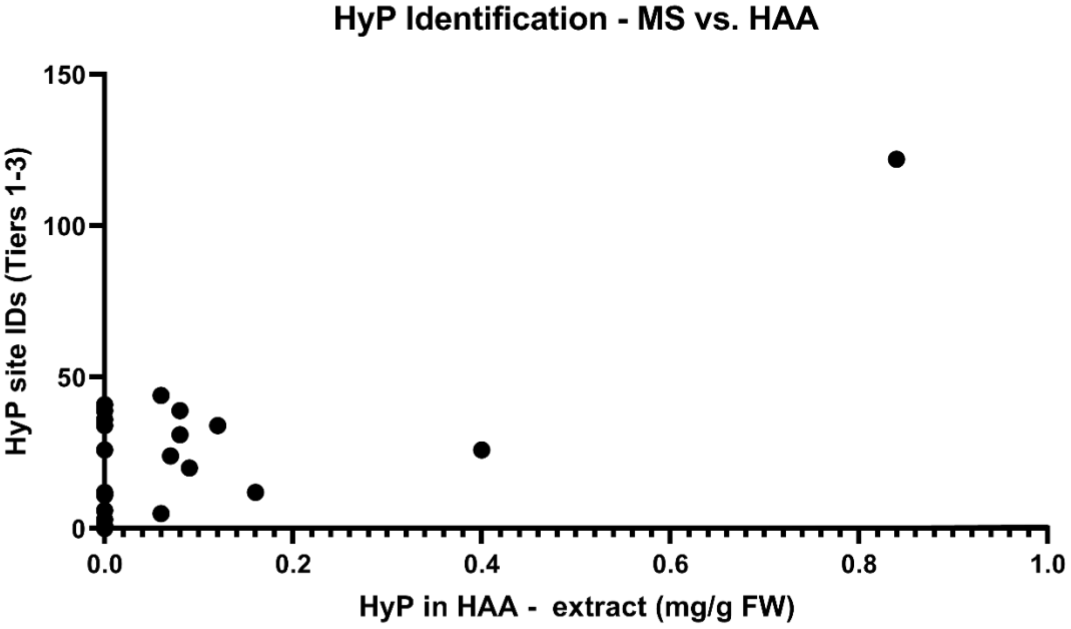


Figure 3.2: HyP identification – MS vs. HAA

The total number of HyP sites identified per sample using MS software analysis are plotted against HyP concentration (mg/g FW) as determined by HAA (extracts). Data points represent each sample which has at least one site in Tiers 1-3. Samples which tested below the limit of quantification for HyP in the HAA (6.25 pmoles/μL) are recorded as “0” for the purposes of this plot as well as correlation and outlier tests.

It is noteworthy that 14 out of 24 total samples in the top three tiers tested below the limit of quantification for HyP in the HAA (6.25 pmol/μL). For the purposes of ranked correlation and outlier analyses, samples which tested below the LOQ for HyP are recorded as “0”. We acknowledge that testing below the LOQ does not define the HyP presence as “0”; however, samples were recorded as such for the feasibility of data assessment. A one-tailed, ranked correlation at a 95% confidence interval was calculated to assess the relationship between HyP detected in HAA of sample extracts and the number of HyP sites identified by MS software analysis. A weak positive correlation was observed, r(22) = 0.32, p = .066. The ROUT method identified peanut as an outlier (Q = 1%) and, when excluded, the Spearman r value decreased to r = 0.210. This decrease, observed in a Spearman’s ranked correlation which mitigates the effect of outliers on the correlation coefficient, indicates the high degree of disarray across the majority of datapoints, suggesting that HAA is not a good indicator of the number of HyP sites identifiable by MS. This is likely due to the inability of HAA to detect HyP at concentrations ≤ 6.25 pmol/μL. While the MS methods used in these experiments do not function to quantify HyP, they are able to identify residues in the primary structure. This indicates which samples do contain HyP even across those which tested below the LOQ in the HAA.

A linear correlation between HyP identification using HAA – and MS-based methods would indicate HyP detection using MS would have similar efficacy as HAA. While it appears that there is little relationship between HAA- and MS-based HyP identification, the sample size is relatively small due to the lack of detection in 14 out of 24 samples which have sites classified to the top three tiers. Additionally, because the protein databases vary in reliability between tiers, data are only truly comparable within their respective tiers. Finally, the variables being compared are not the same – one is a mg/g abundance measure while the other is a count of residue sites. However, the lack of a correlation may indicate a difference in efficacy when an MS-identification method is used compared to an HAA method. As shown in Figure (heat map, E), Spearman's ranked correlation analyses, r_s, were performed among the following variables: total database sequences, total peptides identified, HyP residues identified, and proline residues identified. A positive correlation was observed between total HyP residues identified and total peptides identified, $r_s(20) = .79$. $p = 7.236 \times 10^{-6}$ (Figure 3.3-B, E). The greatest Spearman's ranked correlation coefficient was observed between number of proline residues and total peptides identified, $r_s(20) = .99$, $p = 1.783 \times 10^{-17}$, indicating that they are strongly associated(Figures 3.3-E and 3.4).

To test for linearity, a Pearson correlation, r_p, was also calculated. A perfect linear correlation of these variables ($r_p = 1$) would indicate that, on average, each identified peptide contained one proline residue (modified or unmodified). A positive linear correlation was observed between proline residues and total peptides identified, (Figure 3.4). Thus, a $r_p = 0.99$ indicates, on average, nearly one proline residue per peptide observed (Figure 3.4). While not surprising, this relationship does provide relevant context for interpreting sample quantities of HyP sites. It can now be assumed that any relationship observed between the number of HyP sites in a sample to its number of proline residues can also be generally assumed for that between HyP residues identified and total peptides identified. The lowest correlation coefficient was observed between total database sequences and proline residues identified, $r_s(20) = .54$, $p = .004$, indicating a weak positive correlation (Figure 3.3-E).

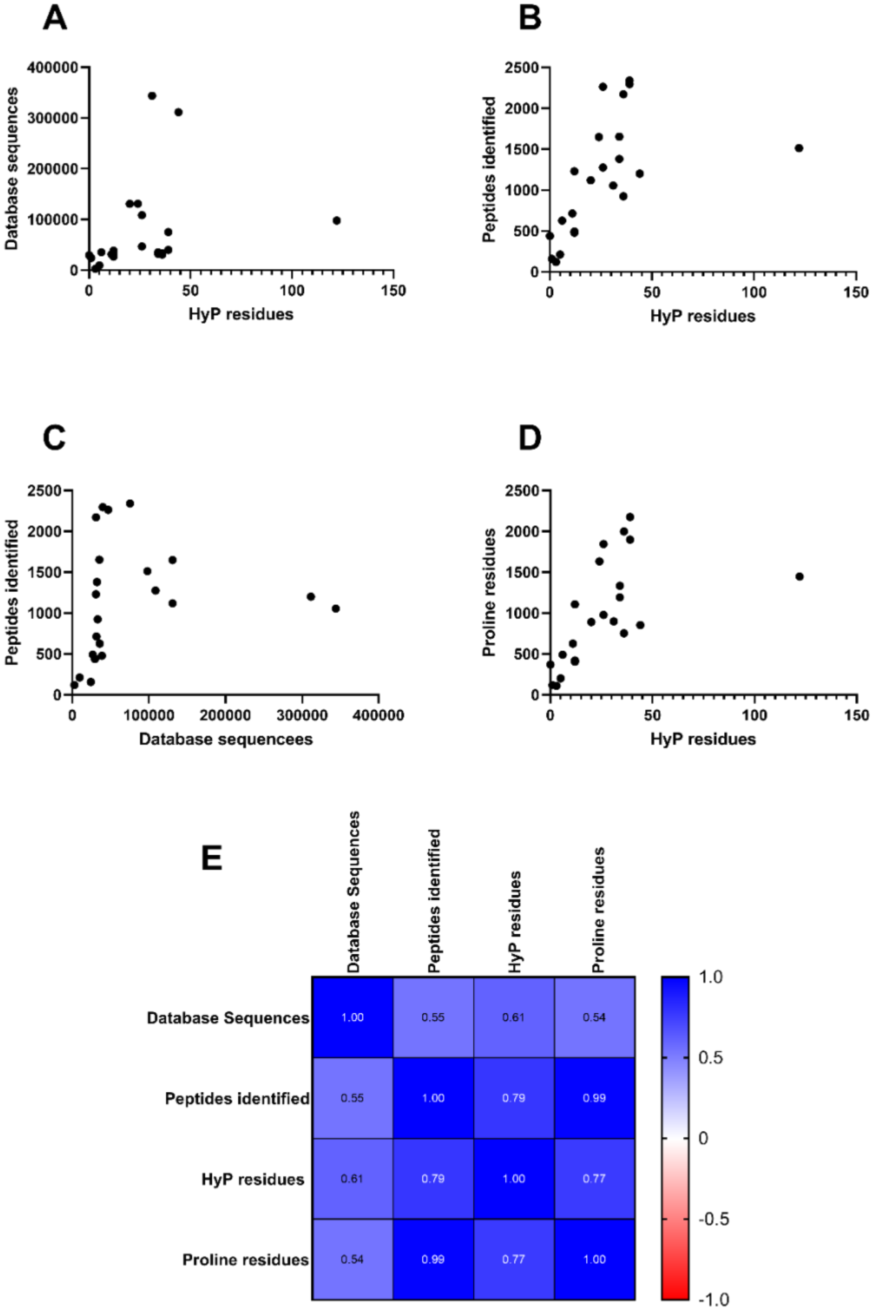


Figure 3.3: Correlation analysis of multiple variables between samples with sites in Tiers 1, 2, and 3

All sites in Tiers 1-3, as gathered from “peptide” PEAKS export, are represented in a multi-variable dot plot analysis to assess the following correlations: A) The number of HyP residues identified is plotted against the total number of sequences in a sample’s respective database; B) The number of HyP residues identified is plotted against the total number of peptides identified; C) The total number of sequences in a respective database is plotted against the total number of identified peptides for that database search; D) The number of HyP residues identified is plotted against the number of proline residues identified in a respective search; E) Spearman r correlation analyses were performed among the following variables: total database sequences *(DB sequences)*, total peptides identified, HyP residues identified, and proline residues identified.

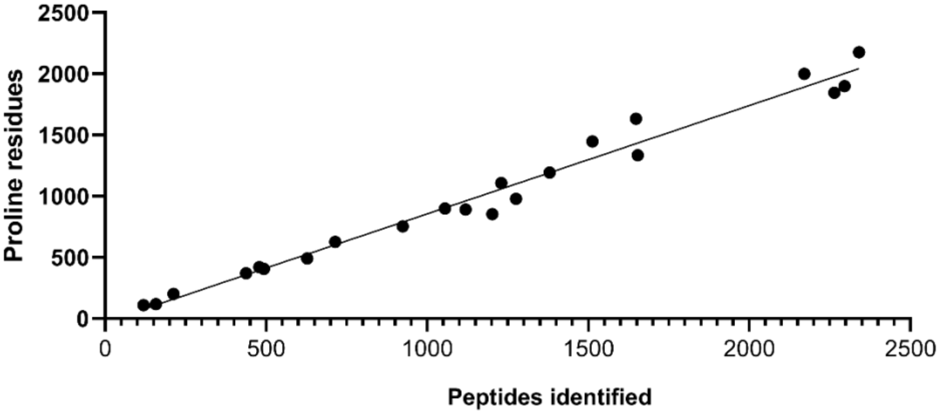


Figure 3.4: Relationship between peptides identified and proline residues

A positive, linear correlation is observed between the number of peptides identified and proline residues among all samples which have sites in Tiers 1-3.

The second-lowest Spearman's correlation value was observed between total database sequences and total peptides identified, $r_s(20) = .55$, $p = .004$, indicating a weak positive correlation comparable to that between database sequences and proline residues 93.3-C, E). Both metrics suggest that database size is only loosely indicative of the number of identifiable proline residues and the number of peptides.

In this case, the relatively weak correlations could be attributed to several factors. The complexity of a plant as well as the seed itself may impact the relationship between database size and the number of peptides or proline residues detected. Proteomic analyses were performed only on the seed of the plant while protein databases typically represent the whole plant. Thus, if a seed contains a relatively small number of proteins compared to the plant as a whole, it is plausible that the number of peptides identified would be low compared to the size of the reference database. Conversely, if a plant's database is largely comprised of proteins which have been sequenced and uploaded from studies of the plant's seeds, then it is understandable that there would be a very high number of peptides identified compared to the total database sequences.

Other factors that may impact the relationship between these variables include efficacy of the extraction buffer and digestion enzymes used. For instance, glutenins and gliadins make up most wheat protein content and are only soluble in alcohol. The urea/thiourea/tris buffer used was not optimal for wheat protein extraction, so it is understandable that the peptide IDs are relatively low compared to the *Triticum aesitivum* database size.

3.3.3 Tier 1 HyP sites normalized to proline residues

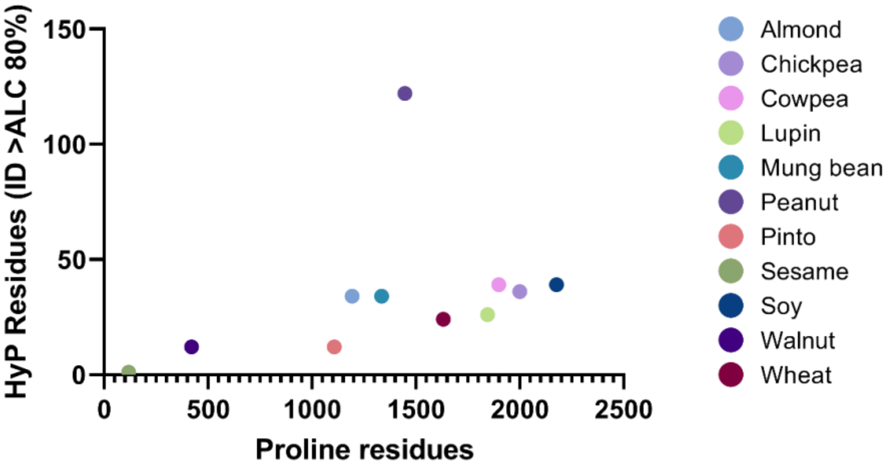


Figure 3.5: Tier 1 – HyP residues per proline residue

HyP sites were plotted against total proline residues identified in peptides from PEAKS DB searches against native reference protein databases (Tier 1). Spearman r correlation analysis was performed to assess correlation between the two variables. The ROUT method was applied to assess outliers, as well as an additional correlation analysis excluding the outlier.

Though only 11 were eligible for Tier 1 comparison, a moderately strong positive Spearman's rank correlation was observed between total proline and total HyP residues,

$r_s(9) = .732$, $p = .007$, among all samples with Tier 1 HyP sites. In the same data pool, ROUT (Q = 1%) identified peanut as an outlier among all samples studied against their native protein databases (Figure 3.5). Though a ranked correlation minimizes the impact outliers have on a dataset, the small size of the Tier 1 sample pool encouraged the assessment of data excluding outliers. When peanut was excluded, the Spearman's correlation of Tier 1 sample HyP sites between proline and HyP residues increased to $r_s(8) = .88$, $p = .001$, indicating a strong positive correlation between total proline residues and total HyP residues in our most reliable data points. Furthermore, when peanut was excluded, a positive linear relationship was observed between total proline and HyP residues in samples with Tier 1 sites, $r_p(8) = .86$, $p = .001$ (Figure 3.5). The linear, positive relationship between total proline and total HyP residues may suggest that an increase in proline is associated with an increase in HyP sites. Furthermore, that peanut is an outlier in this dataset indicates peanut's uniqueness in number of HyP sites compared to other samples of similar confidence, despite an unremarkable proline count.

Identification of peanut as an outlier encourages further investigation of the ways and degree to which peanut is unique from other samples. There are more HyP residues per proline residue in peanut than any other sample by approximately 300% (Table 3.4). This indicates the degree to which peanut's prolyl hydroxylation rate, referring to the proportion of identifiable proline residues which are hydroxylated, exceeds others of a similar confidence level. The peptide.csv export (PEAKS 8.5) allowed for each unique peptide to be identified as either containing- or lacking-HyP. The peptides were sorted to only include those at 80% or greater ALC.

Table 3.4: HyP residues as percentage of proline residues

	PROLINE RESIDUES	HYP RESIDUES	% MODIFIED PROLINES (-OH)
ALMOND	1193	34	2.85%
CHICKPEA	2000	36	1.80%
COWPEA	1899	39	2.05%
LUPIN	1845	26	1.41%
MUNG BEAN	1335	34	2.55%
PEANUT	1447	122	8.43%
PINTO	1107	12	1.08%
SESAME	117	1	0.85%
SOY	2177	39	1.79%
WALNUT	420	12	2.86%
WHEAT	1632	24	1.47%

Tier 1 HyP sites and total proline residues identified in peptides from PEAKS DB searches against native reference or substantial protein databases using software analysis. % Modified prolines was calculated by HyP residues/proline residues expressed as a percentage.

The software compiled peptides suspected to be derived from the same protein and reports the most confident compilation of MS/MS spectra that make up a plausible primary structure of the MS ion. Because the algorithm determines the most confident version of the amino acid sequence for that peptide based on the integrated database and *de novo* analysis, peptides and fragments detected to contain hydroxyproline are not always reported in the sequence. Because the DDA method used in this experiment only fragmented the top 20 most abundant MS ions in a cycle, MS ions which may contain HyP would not be selected if they are of lower abundance. Therefore, the precursor (MS) ion would not be fragmented and would not show fragment ions. Thus, overlapping of fragments to identify primary structure and potential HyP sites would not be possible.

Furthermore, even if the ion was selected for fragmentation during acquisition, *in-silico* overlay may not present lower-confident MS1 spectra and its respective MS2 scans. The lower-confidence peptides with HyP sites may not be displayed even if a HyP residue itself is highly confident. Because we are specifically interested in the HyP sites (rather than the peptides as a whole), it was necessary to design an analysis method to place a greater weight on our specified criteria of interest. For this reason, we used the PSM ions.csv export which presented every spectral event which matched with a theoretical mass from the reference database. This ensures that no mass event would be excluded by the algorithm thus allowing us to manually filter using our pre-defined criteria. This method aims to refine the dataset to only datapoints of greater confidence which fulfill the previously specified criteria of AScore and ALC as well as possessing bilateral fragmentation.

3.3.4 HyP identification by manual analysis

To further investigate HyP site quality and confidence, data was also analyzed by a manual approach. In addition to the confidence parameters provided by PEAKS, we were also interested in fragmentation patterns as a measure of HyP site reliability. By requiring bilateral fragmentation around the HyP site, the number of possible residue combinations for the fragment’s monoisotopic mass decreases. This allows for increased confidence in the combinations that are matched with a theoretical mass from the database.

Manual analyses were conducted in each of the four tiers (Figure 3.6) to determine which sites, of those identified by the software, had bilateral fragmentation.

This was expressed by a metric of 'reduction percentage' to indicate the degree of decrease in total number of HyP sites from software analysis to manual analysis. Reduction percentage was calculated by the following formula: $\% = (\text{Sites}_{\text{software}} - \text{Sites}_{\text{manual}})/(\text{Sites}_{\text{software}}) * 100$. On average and across all tiers, the total number of HyP sites observed per sample decreased by 76.9±18.1% after the manual refining process excluding samples in which HyP sites were neither identified using software nor manual analysis (Table 3.5). A positive reduction represents effective refinement and increased confidence with the manual method across all samples. However, data from different tiers are different in their acquisition and are not truly comparable. Still, this metric can give some indication of the high amount of variability from sample to sample in percent reduction (Table 3.5).

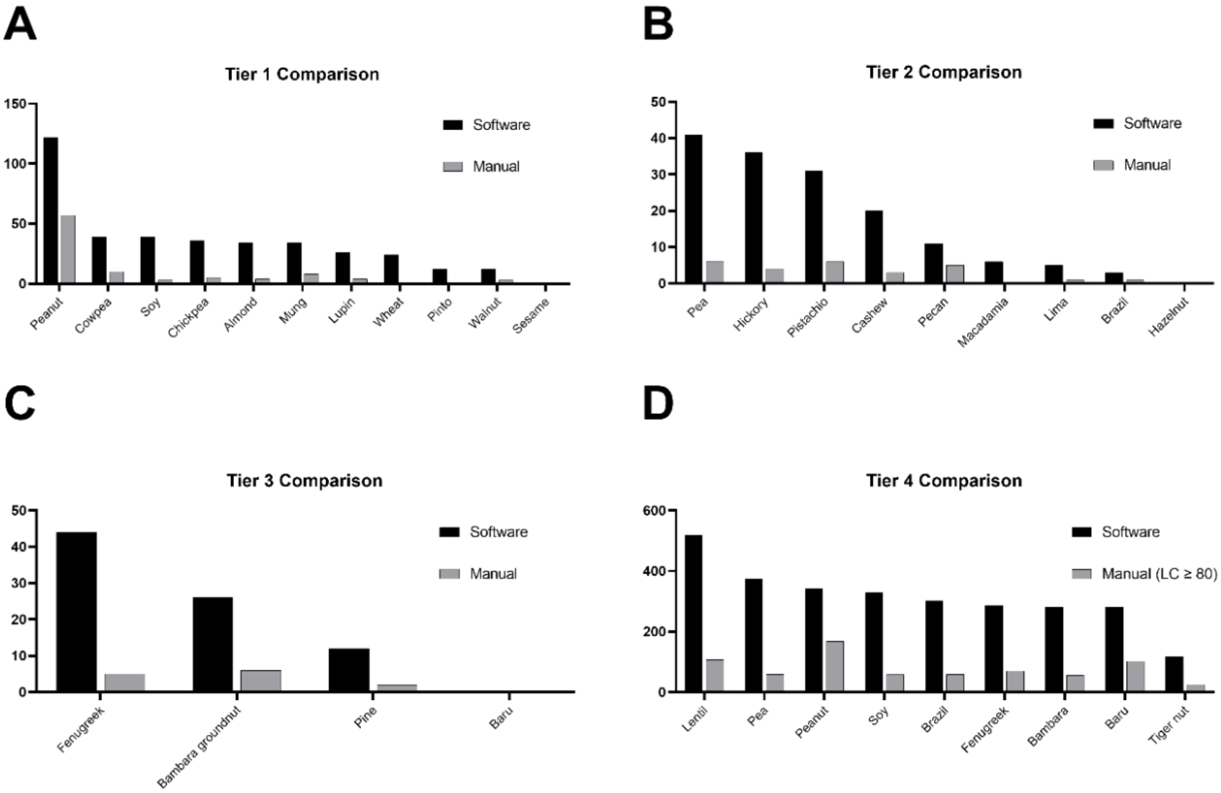


Figure 3.6: HyP residue quantities yielded by software vs. manual analysis methods

Each sample is represented by the number of HyP sites identified using software analysis (black bars) compared to manual analysis in Tier 1 (A), Tier 2 (B), Tier 3 (C), and Tier 4 (D). Tier 4 (D) required a LC ≥ 80% threshold for manual confidence evaluation. A substantial reduction in total site quantity is observed across all four tiers. These results represent successful refinement of sites according to evaluation criteria of bilateral fragmentation, thus increasing confidence of passing sites.

Peanut did display the smallest rates of reduction at 53.3% and 50.9% among all samples in Tiers 1-3 and among all samples in Tier 4, respectively (Figure 3.6, Table 3.5). This indicates that peanut had a greater number of HyP sites that filled the bilateral fragmentation criterion even before it was enforced compared to other samples. Even during software analyses when bilateral fragmentation was not enforced, a greater proportion of these sites did, in fact, present bilateral fragmentation. A greater proportion of HyP sites in peanut were classified as confident when only software data were analyzed.

The only sample that is similar in reduction rate to peanut is pecan, showing a 54.5% reduction in total HyP sites (Table 3.5). This is comparable with peanut, indicating that more of pecan’s HyP residues also have bilateral fragmentation than those of other samples. The lowest rate of reduction was held by sesame at 0%; however, the sample size was only one single HyP residue identified by software analysis. The ROUT method (Q = 5%) was used to identify pecan, sesame, and both database and *de novo* analyses of peanut as outliers. The remaining samples across all tiers (excluding Baru nut and hazelnut which identified zero sites in software and manual analyses) had an average percent reduction of approximately 81.4±9.3%. This indicates that, even when outliers are removed, there is high variability in the reduction percentage when comparing the number of HyP sites identified in manual and software analysis methods.

HyP Site ID Quantities

Tier 1	Software	Manual	% Reduction
Almond	34	4	88.2%
Chickpea	36	5	86.1%
Cowpea	39	10	74.4%
Lupin	26	4	84.6%
Mung bean	34	8	76.5%
Peanut*	122	57	53.3%
Pinto	12	1	91.7%
Sesame	1	1	0.0%
Soy*	39	3	92.3%
Walnut	12	3	75.0%
Wheat	24	1	95.8%

Tier 2	Software	Manual	% Reduction
Cashew	20	3	85.0%
Lima bean	5	1	80.0%
Hazelnut	0	0	-
Hickory	36	4	88.9%
Pistachio	31	6	80.6%
Macadamia nut	6	0	100.0%
Pecan	11	5	54.5%
Brazil nut *	3	1	66.7%
Pea	41	6	85.4%

Tier 3	Software	Manual	% Reduction
Pine nut	12	2	83.3%
Fenugreek*	44	5	88.6%
Bambara groundnut*	26	6	76.9%
Baru*	0	0	-

Tier 4	Software	Manual	% Reduction
Bambara	281	56	80.1%
Brazil	302	59	80.5%
Fenugreek	286	70	75.5%
Tiger nut	116	25	78.4%
Peanut	342	168	50.9%
Soy	329	58	82.4%
Lentil	517	107	79.3%
Pea	374	59	84.2%
Baru	281	102	63.7%

Table 3.5: HyP site refinement from software to manual analysis

Total sites identified using software vs. manual approach are displayed as well as % reduction = (Sites$_{software}$ - Sites$_{manual}$)/(Sites$_{software}$).

3.3.5 *De novo* analyses of HyP-containing peptides

For Tier 4, the same "de novo peptides.csv" PEAKS export was used as with the software analysis; however, MS Excel macros were designed to further customize the analytical method to yield a breakdown of site quantities at different confidence thresholds for each sample. This feature is not currently available using the software alone and was conducted manually.

Table 3.6: Sites identified by *de novo* at increasing minimum local confidence (LC) thresholds

Site Quantities	***Total Sites***	***LC ≥ 50***	***LC ≥ 60***	***LC ≥ 70***	***LC ≥ 80***	***LC ≥ 90***	***LC ≥ 95***
Bambara	281	226	162	105	56	35	18
Fenugreek	286	222	174	123	70	27	10
Peanut	343	294	245	205	168	137	113
Soy	329	261	183	114	58	25	9
Tiger nut	116	98	72	48	25	7	3
Brazil	302	248	183	104	59	41	26
Lentil	517	392	287	184	107	58	28
Baru	281	227	187	148	102	54	30
Pea	374	285	185	112	59	29	10

Percentages	***LC ≥ 50***	***LC ≥ 60***	***LC ≥ 70***	***LC ≥ 80***	***LC ≥ 90***	***LC ≥ 95***
Bambara	80.4%	57.7%	37.4%	19.9%	12.5%	6.4%
Fenugreek	77.6%	60.8%	43.0%	24.5%	9.4%	3.5%
Peanut	85.7%	71.4%	59.8%	49.0%	39.9%	32.9%
Soy	79.3%	55.6%	34.7%	17.6%	7.6%	2.7%
Tiger nut	84.5%	62.1%	41.4%	21.6%	6.0%	2.6%
Brazil	82.1%	60.6%	34.4%	19.5%	13.6%	8.6%
Lentil	75.8%	55.5%	35.6%	20.7%	11.2%	5.4%
Baru	80.8%	66.5%	52.7%	36.3%	19.2%	10.7%
Pea	76.2%	49.5%	29.9%	15.8%	7.8%	2.7%

These data provide information for samples which lack any type of database. *De novo* sequencing was performed on some samples in addition to Tier 1 analysis to allow for bias evaluation of *de novo* methods. Reference proteomes exist for peanut and soy;

however, *de novo* sequencing yielded similarly high results in peanut as compared to PEAKS DB and manual analyses. Peanut does not have the most initial HyP sites identified by *de novo* sequencing. Instead, it is only the second highest in total HyP sites among the seven total samples evaluated in Tier 4 (Table 3.6). However, peanut does have more high-confidence sites than any other sample. Other samples have high identification rates but the confidence that those sites are truly occurring is lower than that of peanut. Indeed, nearly 33% of peanut's total sites have local confidence scores of 95 or higher. The next highest sample, Baru nut, shows an LC score of ≥ 95 in 10.7% of its sites (Table 3.6).

This evaluation method adds value to *de novo* sequencing because, even though there is no database against which m/z values can be cross-referenced, there is a way to evaluate single residues even if the average confidence of the peptide is slightly lower. It is true that, for HyP IDs to be meaningful in *de novo* data, they must be part of a relatively confident peptide so the protein to which it may belong could be deduced. However, as long as a reasonable confidence threshold for the total peptide sequence (i.e., ALC ≥ 80%) is maintained, it is valuable to have the ability to put an even greater pinpoint focus on the residues of interest – in this case, hydroxyproline.

3.3.6 HyP sites across protein and sample types

Further meaning can be contributed to identified HyP sites by also identifying the protein to which the modified amino acid belongs. This identification was performed on all HyP sites which were classified in Tiers 1-3. Sites in Tier 4 were identified exclusively by *de novo* sequencing and, by definition, do not have a protein database for reference.

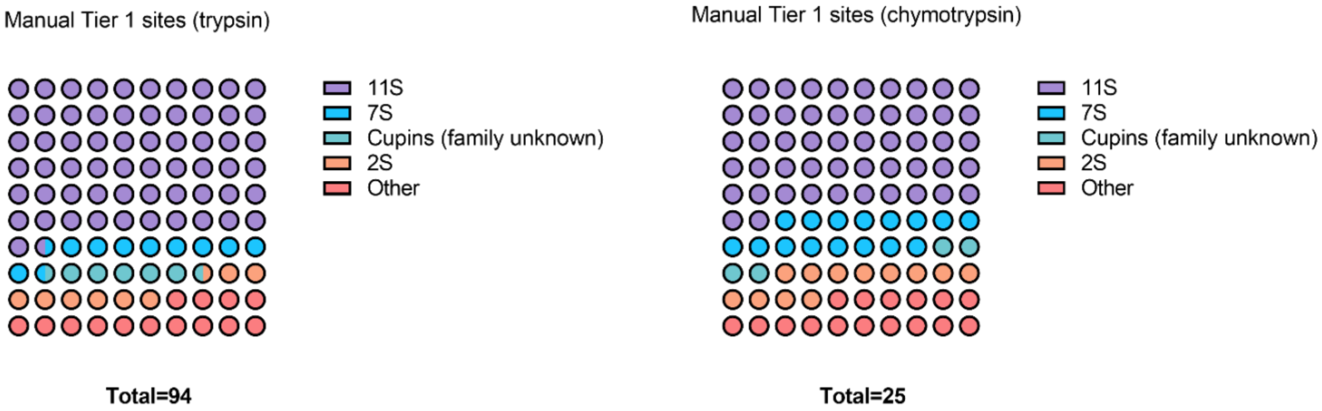


Figure 3.7: HyP sites by protein classification

Each of the most confident, robust HyP sites is represented (i.e., Tier 1 HyP sites after manual refinement). Tryptic and chymotryptic peptides yielded 94 and 25 total HyP sites, respectively, in the top tier after manual analysis. The proteomic location of each site was identified and classified by protein type. Major seed storage proteins represented the majority of HyP site locations with "Other" encompassing a variety of cytosolic and membrane-bound proteins.

In both tryptic and chymotryptic peptides, most HyP sites are identified on 11S globulins (Figure 3.7). Proteins are defined by the solvent in which they are the most soluble. This may indicate that 11S proteins were preferentially extracted. Conversely, it could mean that a higher proportion of HyP sites occur on 11S proteins than other seed storage proteins. Because these are Tier 1 HyP sites, the databases are the most reliable of any others in the study. Still, because protein sequences are largely obtained by genome translation, it is possible that some sequences could be missing from proteome.

3.3.7 Implications of HyP for allergy

The findings of this study indicate an unremarkable number of HyP sites in most samples, including but not limited to allergenic tree nuts such as hazelnut, walnut, and pecan. Some pecan and walnut proteins share high levels of homology[118] and cross-reactivity between the two is common[119]. Many times, co-allergy to both pecan and walnut is observed[120]. Although lower than that between walnut and pecan, serum IgE cross-reactivity has also been observed between walnut- and hazelnut-allergic individuals[119,121].

Hazelnut, walnut, and pecan are all major tree nut allergens which can present cross reactivity among each other; however, each of them presented low levels of HyP by the methods outlined in this thesis. No HyP sites were identified in hazelnut using either MS analysis method or HAA. Pecan did present a similar proportion of confident HyP sites to total HyP sites compared to peanut; however, peanut displayed a substantially higher number of total sites. The lack of HyP identification in a frequent allergen such as hazelnut indicates that HyP is not a prerequisite for plant allergenicity. HyP was not identified by HAA of soluble proteins in walnut, and walnut was second lowest in HyP site identification using MS software analysis. The low levels of identification in potent tree nut allergens, pecan and walnut, allow for speculation that HyP levels may not be associated with allergenicity.

However, in both PEAKS DB and *de novo* analyses, peanut presents as a clear outlier in total HyP abundance, total HyP sites, and total HyP residues per proline residue among comparable samples. Further study is required to determine the location of these HyP sites with regard to known peanut allergens and IgE epitopes, as well as impact of

HyP on IgE binding levels and immunoreactivity. However, the levels of HyP in peanut seed storage proteins found in this study, when considered with the immunogenic importance of the linear HyP-containing motifs in Ara h 2 and Phl p 1 suggest that there may be a connection between HyP and some aspects of allergy.

If a connection between HyP and food allergies is determined in the future, HyP would be an important factor to consider when evaluating novel foods for potential allergenicity and risk to the population in the future.

3.4 Conclusions

In these experiments, HyP sites were identified in 25 out of 26 total samples using MS analysis. These data indicate that MS can identify HyP in samples which HAA cannot. While a total abundance exceeding the LOQ is required for HAA, HyP sites need only to occur on a relatively abundant peptide to be detected by MS. Additionally, software analyses can indicate the primary structure of peptides in which HyP is found and position of putative sites within those peptides. Furthermore, though site identification by software alone is helpful, it can be customized using bespoke evaluation criteria by pairing it with manual fragmentation analysis. This allows a user to identify highly confident HyP sites according to differently emphasized criteria than that provided by PEAKS's PTM Profile function. Finally, even when a database is unavailable, preliminary data for hydroxyproline site identification can be collected using *de novo* sequencing. Though it is less reliable than data acquired using a database, it can still indicate HyP position and surrounding residues as well as a starting point for further investigation.

HyP site position and surrounding amino acid sequence can provide meaningful context when interpreting HyP in the context of food allergens, specifically in determining HyP-containing amino acid motifs. Currently, only a couple of motifs have been identified; however, future motif studies across Viridiplantae could potentially reveal patterns of prolyl hydroxylation which would clarify the mechanism of the modification and implications of HyP in allergy. HyP sites associated with seed storage proteins may serve as subjects for further investigation in their potential role in immunoreactivity. These investigations could expand upon the known location of highly abundant HyP on 2S albumin and potent peanut allergen, Ara h 2, and could provide further information on potential significance of HyP presence in seed storage proteins.

REFERENCES

(1) Petersen, A.; Schramm, G.; Schlaak, M.; Becker, W. M. Post-translational modifications influence IgE reactivity to the major allergen Phl p 1 of timothy grass pollen. *Clinical and experimental allergy : journal of the British Society for Allergy and Clinical Immunology* **1998,** *28* (3), 315.

(2) Petersen, A.; Suck, R.; Hagen, S.; Cromwell, O.; Fiebig, H.; Becker, W.-M. Group 13 grass allergens: Structural variability between different grass species and analysis of proteolytic stability. *Journal of Allergy and Clinical Immunology* **2001,** *107* (5), 856.

(3) Koppelman, S. J.; Wensing, M.; Ertmann, M.; Knulst, A. C.; Knol, E. F. Relevance of Ara h1, Ara h2 and Ara h3 in peanut-allergic patients, as determined by immunoglobulin E Western blotting, basophil–histamine release and intracutaneous testing: Ara h2 is the most important peanut allergen. **2004,** *34* (4), 583.

(4) Lehmann, K.; Schweimer, K.; Reese, G.; Randow, S.; Suhr, M.; Becker, W. M.; Vieths, S.; Rösch, P. Structure and stability of 2S albumin-type peanut allergens: implications for the severity of peanut allergic reactions. *The Biochemical journal* **2006,** *395* (3), 463.

(5) Chatel, J. M.; Bernard, H.; Orson, F. M. Isolation and characterization of two complete Ara h 2 isoforms cDNA. *International archives of allergy and immunology* **2003,** *131* (1), 14.

(6) Bernard, H.; Guillon, B.; Drumare, M. F.; Paty, E.; Dreskin, S. C.; Wal, J. M.; Adel-Patient, K.; Hazebrouck, S. Allergenicity of peanut component Ara h 2: Contribution of conformational versus linear hydroxyproline-containing epitopes. *The Journal of allergy and clinical immunology* **2015,** *135* (5), 1267.

(7) Hazebrouck, S.; Guillon, B.; Paty, E.; Dreskin, S. C.; Adel-Patient, K.; Bernard, H. Variable IgE cross-reactivity between peanut 2S-albumins: The case for measuring IgE to both Ara h 2 and Ara h 6. *Clinical and Experimental Allergy* **2019,** *49* (8), 1107.

(8) Camerini, S.; Mauri, P. The role of protein and peptide separation before mass spectrometry analysis in clinical proteomics. *Journal of Chromatography A* **2015,** *1381*, 1.

(9) Gorres, K. L.; Raines, R. T. Prolyl 4-hydroxylase. *Critical Reviews in Biochemistry and Molecular Biology* **2010,** *45* (2), 106.

(10) Han, X.; He, L.; Xin, L.; Shan, B.; Ma, B. PeaksPTM: Mass spectrometry-based identification of peptides with unspecified modifications. *Journal of proteome research* **2011,** *10* (7), 2930.

(11) Stanke, M.; Steinkamp, R.; Waack, S.; Morgenstern, B. AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic acids research* **2004,** *32* (Web Server issue), W309.

(12) Stanke, M.; Keller, O.; Gunduz, I.; Hayes, A.; Waack, S.; Morgenstern, B. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic acids research* **2006,** *34* (Web Server issue), W435.

(13) Chen, S.; Downs, M. L. Proteomic Analysis of Oil-Roasted Cashews Using a Customized Allergen-Focused Protein Database. *Journal of proteome research* **2022,** *21* (7), 1694.

(14) FAIRsharing.org: HWG, 2016, DOI: 10.25504/FAIRsharing.srgkaf 10.25504/FAIRsharing.srgkaf.

(15) Geiselhart, S.; Hoffmann-Sommergruber, K.; Bublin, M. Tree nut allergens. *Molecular Immunology* **2018,** *100*, 71.

(16) Goetz, D. W.; Whisman, B. A.; Goetz, A. D. Cross-reactivity among edible nuts: double immunodiffusion, crossed immunoelectrophoresis, and human specific IgE serologic surveys. *Annals of Allergy, Asthma & Immunology* **2005,** *95* (1), 45.

(17) Andorf, S.; Borres, M. P.; Block, W.; Tupa, D.; Bollyky, J. B.; Sampath, V.; Elizur, A.; Lidholm, J.; Jones, J. E.; Galli, S. J.et al. Association of Clinical Reactivity with Sensitization to Allergen Components in Multifood-Allergic Children. *The Journal of Allergy and Clinical Immunology: In Practice* **2017,** *5* (5), 1325.

(18) Asero, R.; Mistrello, G.; Roncarolo, D.; Amato, S. Walnut-induced anaphylaxis with cross-reactivity to hazelnut and Brazil nut. *Journal of Allergy and Clinical Immunology* **2004,** *113* (2), 358.

CHAPTER 4: SUMMARY, CONCLUSIONS, AND FUTURE WORK FOR HYDROXYPROLINE IN PLANT SEED PROTEINS

4.1 Introduction

Posttranslational modifications can impact the function and role of a protein and are the cause for substantial biological research[122-125]. Liquid chromatography tandem mass spectrometry (LC-MS/MS) is currently the most advanced tool for posttranslational modification (PTM) identification in plant proteins, especially when searching with relation to plant allergens. One PTM of interest to allergy is hydroxyproline (HyP), for which there is evidence of its impact on IgE binding capacity and mediator release[68,70,110]

Characterization of the amino acid sequence surrounding a putative HyP site can provide valuable context for the protein on which the site occurs, as well as increase confidence in the position and identity of the HyP site itself.

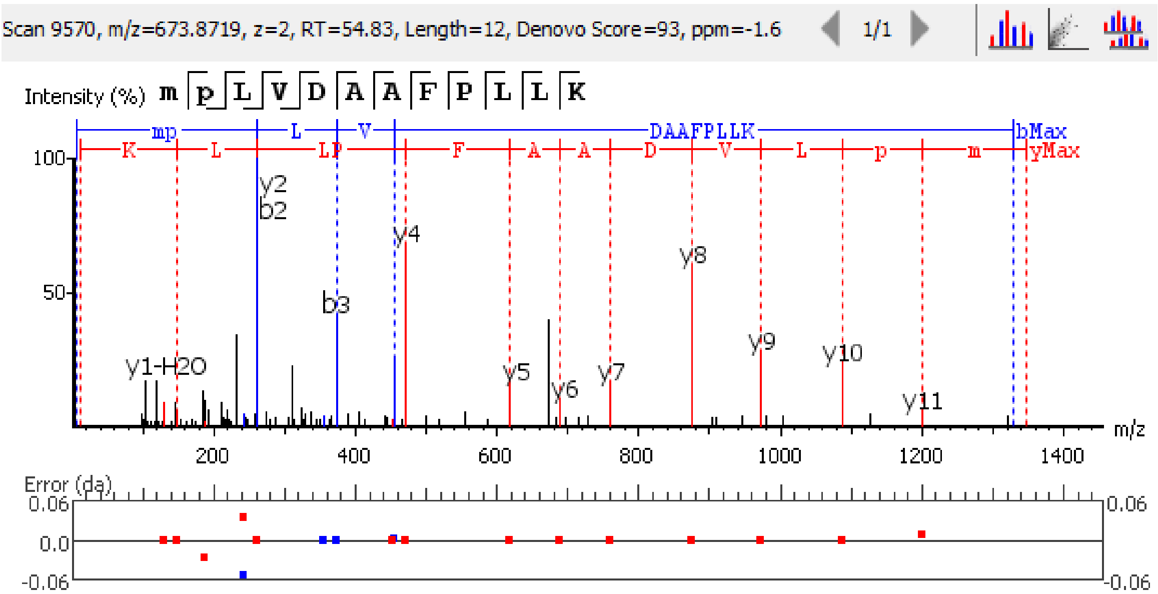
4.2 Limitations and considerations of the study

Protein extractability and database availability are two significant challenges in this study. First, we acknowledge that each protein type extracts differently in a chaotropic buffer and that each sample has different distributions of these proteins. To limit this variability, optimization of buffers for each sample would be required. Even with optimized buffers, 100% protein recovery would not occur thus still incurring some level of bias. Secondly, database availability substantially impacts the quality and comprehensiveness of MS data. To mitigate this, databases were curated from a variety of sources to produce as exhaustive a proteome as possible.

Additionally, evaluating HyP sites by tier functioned to address some of the variability in database reliability. Still, we acknowledge that each species is unique in its total protein count as well as distribution of proteins in the seed compared to the rest of the plant. Thus, it is expected that databases were not all-encompassing, a factor that should be considered when drawing conclusions.

In addition to limitations in sample preparation and database availability, there are limitations related to the software's interpretation of raw data. Amino acids leucine (Leu) and isoleucine (Ile) are identical in molecular mass and are therefore largely indistinguishable during analysis[126]. If a sample is being cross-referenced to a protein database, this ambiguity can be mitigated by the known amino acid sequences supplied by the database. However, if the sample is being sequenced by *de novo* or if it is a portion of the proteome which does not have protein sequences available, incorrect identification of Leu and/or Ile should be considered. This could impact protein identification if sequences containing HyP sites undergo homological searches for protein identification. These searches can provide context to peptides identified by LC-MS/MS by matching them with known sequences from potentially related organisms with varying degrees of confidence. The potential misidentification of Leu and/or Ile should be considered when interpreting these search results.

Methionine oxidation, another common PTM, would add the same mass (+15.99) as HyP; therefore, one must be cognizant of methionine residues in the same MS2 fragments as detected HyP sites. The same mass difference would occur and may not be accurately reflected in position. This decreased confidence should be reflected in position scores like PTM Ascore and local confidence (LC) score to varying degrees based on number of possible amino acid combinations for a given fragment.



Scan 9570, m/z=673.8719, z=2, RT=54.83, Length=12, Denovo Score=93, ppm=-1.6

Figure 4.1: MS1 spectrum for one peptide from lentil digested with trypsin

Complete fragmentation is displayed for full MS2 coverage. Adjacent modified residues of oxidized methionine (m) and hydroxylated proline (p) at the N-terminus of the peptide are confirmed by complete MS/MS fragmentation.

Interestingly, lentil data displayed an instance in which methionine oxidation and proline hydroxylation occurred adjacently (Figure 4.1). As seen in the figure below, this peptide analyzed by *de novo* sequencing displays complete fragmentation thus full MS2 spectra which allows for very high confidence of each residue position and identity. Oxidized methionine shows an LC score of 92 and the HyP residue a score of 94, with an ALC of 93% sequence confidence for the peptide overall. This indicates that methionine oxidation does not prevent hydroxylation of adjacent proline residues. Furthermore, this spectrum exemplifies a strategy for distinction of hydroxyl addition to a residue, showing complete MS2 fragmentation of the proline and methionine residues. This provides high confidence of the position at which the modification occurs.

Limitations of interpreting both Leu and Ile identification as well as differentiating methionine oxidation and prolyl hydroxylation could present challenges in related future work. They may impact accuracy of *de novo* HyP site identification. Additionally, these limitations would likely present a challenge in determining potential motifs associated with HyP sites or signal sequences for this modification and additional steps should be taken to mitigate these challenges.

4.3 Indications of findings

As indicated by hydrolyzed amino acid analysis (HAA), proline and HyP abundance are not correlated in ground sample; however, the two are loosely correlated in soluble proteins. Peanut and Bambara groundnut were the only two outliers in soluble HyP per proline residue among all 26 samples. The Bambara groundnut lacks a native reference proteome or genome-predicted protein database, but peanut displayed the highest number of proline sites during both software and manual analyses.

Peanut was also analyzed using *de novo* sequencing to provide context for other samples with sites in Tier 4. Out of all sites evaluated in peptides at ALC ≥ 80%, peanut had the highest proportion of confident residues and had 58% more confident sites (LC ≥ 80%) than the next-highest sample. Peanut had the greatest number of HyP sites identified by *de novo* sequencing after both software and manual analysis. Because of the established increased IgE binding and immunogenicity associated with HyP[68,70,110] as well as the uniqueness in number of HyP sites identified in the primary structure of major peanut allergens, it seems plausible to infer a connection between the two.

A great increase in reliability and robustness of the data would occur if all MS outputs were run against Tier 1 databases. Currently, only 11 samples have native reference or substantial proteomes which landed them in the top tier. This is a substantial limitation because *de novo* sequencing relies solely on fragment ion overlap and theoretical amino acid masses, along with those of potential PTMs. This produces a lower degree of specificity and confidence in the resulting amino acid sequences.

4.4 Speculatory discussion & further research

While the results of this study provide some interesting indications, several future studies would provide valuable context to the current results. The research presented in this thesis is meant to serve as a starting point for further exploration into the mechanism of food allergy and factors which contribute to elicitation of an allergic response. Here, we explore the PTM profile regarding hydroxyproline in foods. Other PTMs could be further studied to inquire about their potential contribution to food allergies or their prevalence in highly allergenic foods.

Furthermore, it would be interesting to do a bioinformatic study of the motifs that could contribute to proline hydroxylation. The signal to induce prolyl hydroxylation in soluble plant cell proteins is not currently known. Questions still to be answered include what recruits the enzyme, additional functions of prolyl hydroxylation, and mechanism for modification. It is well understood that, in collagen, the HyP contributes structural flexibility to increase stability of a collagen triple helix. In seed storage proteins, however, it appears that the pattern of prolyl hydroxylation is far more random. Thus, it would be beneficial to do a bioinformatical study of various amino acid residue windows surrounding the reliable HyP sites found in this study to investigate whether there is some sort of pattern in amino acid identity, hydrophobicity, acidity, or other molecular property that may point to a pattern for recruitment of prolyl hydroxylase.

It is valuable to increase the universal understanding of the mechanism of food allergy or some molecules that contribute to or are associated with allergenic reaction elicitation. This information could then be considered when 1.) assessing a novel food for potential allergenicity or 2.) exploring the molecular pathway of certain foods which elicit frequent allergic response.

4.5 Conclusion

From the data presented in this thesis, it is now clear that species across Viridiplantae possess the machinery to perform prolyl hydroxylation, a statement previously not known to be true. Additionally, it is clear that peanut is unique in its rate of prolyl hydroxylation compared to other plants with similarly robust databases. These discoveries reveal further research questions about HyP and its relation to allergy that can be explored in the future.

REFERENCES

(1) Duan, G.; Walther, D. The Roles of Post-translational Modifications in the Context of Protein Interaction Networks. *PLOS Computational Biology* **2015,** *11* (2), e1004049.

(2) STULEMEIJER, I. J. E.; JOOSTEN, M. H. A. J. Post-translational modification of host proteins in pathogen-triggered defence signalling in plants. **2008,** *9* (4), 545.

(3) Bond, A. E.; Row, P. E.; Dudley, E. Post-translation modification of proteins; methodologies and applications in plant sciences. *Phytochemistry* **2011,** *72* (10), 975.

(4) Dai, Z.; Hooker, B. S.; Quesenberry, R. D.; Thomas, S. R. Optimization of Acidothermus cellulolyticus Endoglucanase (E1) Production in Transgenic Tobacco Plants by Transcriptional, Post-transcription and Post-translational Modification. *Transgenic Research* **2005,** *14* (5), 627.

(5) Petersen, A.; Schramm, G.; Schlaak, M.; Becker, W. M. Post-translational modifications influence IgE reactivity to the major allergen Phl p 1 of timothy grass pollen. *Clinical and experimental allergy : journal of the British Society for Allergy and Clinical Immunology* **1998,** *28* (3), 315.

(6) Bernard, H.; Guillon, B.; Drumare, M. F.; Paty, E.; Dreskin, S. C.; Wal, J. M.; Adel-Patient, K.; Hazebrouck, S. Allergenicity of peanut component Ara h 2: Contribution of conformational versus linear hydroxyproline-containing epitopes. *The Journal of allergy and clinical immunology* **2015,** *135* (5), 1267.

(7) Üzülmez, Ö.; Kalic, T.; Mayr, V.; Lengger, N.; Tscheppe, A.; Radauer, C.; Hafner, C.; Hemmer, W.; Breiteneder, H. The Major Peanut Allergen Ara h 2 Produced in Nicotiana benthamiana Contains Hydroxyprolines and Is a Viable Alternative to the E. Coli Product in Allergy Diagnosis. **2021,** *12*.

(8) Xiao, Y.; Vecchi, M. M.; Wen, D. Distinguishing between Leucine and Isoleucine by Integrated LC–MS Analysis Using an Orbitrap Fusion Mass Spectrometer. *Analytical Chemistry* **2016,** *88* (21), 10757.