

SPECIAL GUEST EDITOR SECTION

Selection of Tree Nut Allergen Peptide Markers: A Need for Improved Protein Sequence Databases

WEILI XIONG, MELINDA A. MCFARLAND, CARY PIRONE, and CHRISTINE H. PARKER¹

U.S. Food and Drug Administration, Center for Food Safety and Applied Nutrition, 5001 Campus Dr, College Park, MD 20740

Background: To effectively safeguard the food-allergic population and support compliance with food-labeling regulations, the food industry and regulatory agencies require reliable methods for food allergen detection and quantification. MS-based detection of food allergens relies on the systematic identification of robust and selective target peptide markers. The selection of proteotypic peptide markers, however, relies on the availability of high-quality protein sequence information, a bottleneck for the analysis of many plant-based proteomes. **Method:** In this work, data were compiled for reference tree nut ingredients and evaluated using a parsimony-driven global proteomics workflow. **Results:** The utility of supplementing existing incomplete protein sequence databases with translated genomic sequencing data was evaluated for English walnut and provided enhanced selection of candidate peptide markers and differentiation between closely related species. **Highlights:** Future improvements of protein databases and release of genomics-derived sequences are expected to facilitate the development of robust and harmonized LC–tandem MS-based methods for food allergen detection.

Food allergies are a growing public health concern that affect an estimated 3–4% of the total population in the United States (1). Eight allergenic foods, including milk, eggs, fish, crustaceans, tree nuts, peanuts, wheat, and soybean account for the majority of food-triggered adverse immune responses. Tree nuts, as a group, pose severe health problems to allergic individuals (2). Allergic reactions to tree nuts can be potentially life-threatening and, in most cases, tend to be persistent and lifelong (3, 4). A majority of individuals

allergic to tree nuts are allergic to more than one type of tree nut but not necessarily to all tree nuts (5). Currently, there is no curative therapy for tree nut allergies, relying on management by strict allergen avoidance and prompt treatment of allergic symptoms (2). To enhance food safety for allergic consumers, the development of reliable analytical methods for tree nut detection represents a priority for food industry and regulatory agencies.

Immunochemical techniques such as ELISA have been widely used for food allergen detection because of their direct detection of allergen proteins with sufficient sensitivity and standardization for routine food analysis (6). However, the performance of commercial ELISA kits can be markedly affected by types of food processing, sample preparation procedures, selectivity of the antibody, and reference materials used for validation (7–9). Recently, MS-based approaches have emerged as a confirmatory technique for the orthogonal detection and quantification of food allergens. Targeted LC–tandem MS (MS/MS) methods allow multiplexed allergen detection and quantification in a single analytical experiment offering high sensitivity and specificity (10). Despite elevated cost and requirements of specialized expertise for targeted MS assay development, MS methods are more amenable to the analysis of thermally processed foods and less susceptible to cross-reactivity, providing unambiguous protein and peptide confirmation (11). Recent studies reported the development of MS methods for multiple tree nut detection (12–15). Moving forward in the development of robust LC-MS/MS methods for tree nuts, the selection of species-specific tree nut allergen peptide markers requires consideration to the availability of high-quality protein sequences and empirical evaluation of closely related reference food ingredients.

Tree Nut Allergens

Tree nuts as defined by the Food Allergen Labeling and Consumer Protection Act of 2004 include almond, Brazil nut, cashew, chestnut, coconut, hazelnut, pecan, pine nut, pistachio, walnut, and other infrequently or regionally consumed tree nuts. Tree nuts are produced by seed plants from all major lineages, including monocots (e.g., coconut), eudicots (e.g., macadamia nut, almond, and Brazil nut), and gymnosperms (e.g., pine nut). Homologous sequences of allergenic proteins are highly variable across lineages (16). However, sequences of closely related tree nuts from the same family or order often share a high degree of sequence similarity (Figure 1). Commonly consumed tree nuts (e.g., hazelnut, chestnut, pecan, and walnut) are found in the order Fagales. Pecan and walnut both belong to the Juglandaceae family, while chestnut and hazelnut belong to the Fagaceae and Betulaceae families, respectively. Additionally,

Guest edited as a special report on “Mass Spectrometry: Status Quo in Food Allergen and Food Authenticity Applications” by Bert Popping and Carmen Diaz-Amigo.

This research was funded by the U.S. Food and Drug Administration (FDA), Center for Food Safety and Applied Nutrition. Weili Xiong and Cary Pirone were supported, in part, by an appointment to the Research Participation Program, administered by the Oak Ridge Institute for Science and Education (ORISE) through an interagency agreement between the U.S. Department of Energy and the FDA.

Supplemental information and color images are available on the *J. AOAC Int.* website, <http://aoac.publisher.ingentaconnect.com/content/aoac/jaoac>

¹Corresponding author's e-mail: christine.parker@fda.hhs.gov

DOI: <https://doi.org/10.5740/jaoacint.19-0054>

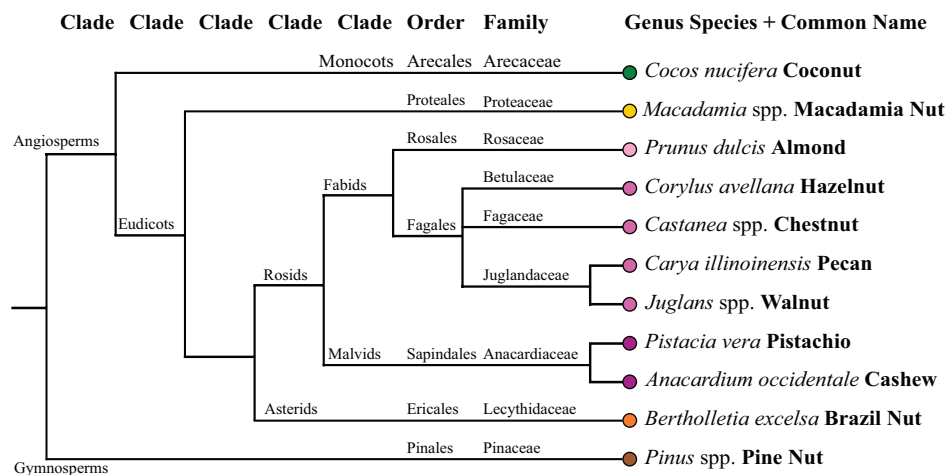


Figure 1. Cladogram showing phylogenetic relationships of selected tree nuts. The unscaled tree was built according to the Angiosperm Phylogeny Group (APG IV 2016). Species of selected tree nuts is in accordance with Section 201(qq) of the Food Allergen Labeling and Consumer Protection Act of 2004.

cashew and pistachio belong to the Anacardiaceae family in the order Sapindales. Despite taxonomic differences, clinical cross-reactions can occur among some tree nuts and legume seeds, which may be due to structural similarities in plant allergen protein families (17).

Plant-derived food allergens can be primarily classified into three protein classes (18): the prolamin superfamily (e.g., 2S albumins and nonspecific lipid transfer proteins), the cupin superfamily (e.g., vicilin-type 7S globulins and legumin-type 11S globulins), and the Bet v 1 superfamily (e.g., class 10 pathogenesis-related proteins). Among these families, seed storage proteins, including 2S albumins, 7S globulins, and 11S globulins, represent the most abundant proteins in tree nuts and legume seeds and are most commonly associated with severe allergic reactivity (19). Additional recognized plant allergens include the profilins and the oleosins, which function as structural proteins (20).

Table 1 lists the tree nut allergens for each of the dominant protein classes, to date, recognized by the World Health Organization (WHO) and International Union of Immunological Societies (IUIS) Allergen Nomenclature Subcommittee (www.allergen.org). Although many tree nut allergen proteins have been identified because of their high abundance and clinical significance, a comprehensive catalog of all tree nut allergens is still in development. Structural homology of allergens within each protein class provides a predictive tool with which to identify potential protein allergens among closely related species. Previous studies have implemented bioinformatic approaches to illustrate sequential and structural homology among allergens belonging to the same superfamily. For example, 2S albumins from walnut (Jug r 1) and pecan (Car i 1), 7S globulins from walnut (Jug r 2) and pecan (Car i 2), and 11S globulins from walnut (Jug r 4) and pecan (Car i 4) have 88, 92, and 95% sequence identity, respectively (21). In addition, high sequence identity (>70%) was observed between allergens in cashew and pistachio (Ana o 1-Pis v 3; Ana o 2-Pis v 5), hazelnut and walnut (Cor a 11-Jug r 6; Cor a 9-Jug r 4), and hazelnut and pecan (Cor a 9-Car i 4). Other studies have shown structural homology among Immunoglobulin E-binding epitopes of legumins and vicilins from peanut and tree nuts (22, 23).

As new tree nut allergens are added to established protein classes, protein homology provides a framework to understand structural and functional relationships between allergenic proteins that influence immunochemical- and clinical-based cross-reactivities among different plant sources (24). Knowledge of structural protein allergen classes and sequence homology in closely related species is critical to the selection of robust, highly specific allergen peptide markers and to the development of targeted MS methods for food allergen analysis.

Considerations for MS Allergen Method Development: Tree Nuts

The development of targeted LC-MS/MS methods for food allergen analysis begins with the selection of peptide markers. Representative peptides, typically ≤ 5 peptides per protein, are empirically selected to function as a proxy for protein quantification across a wide range of food matrices and food processing conditions. The workflow includes a discovery-based proteomic evaluation of candidate peptide markers using full-scan (spectral feature) and MS/MS (spectral counting) methodologies in raw and thermally processed allergen ingredients, cultivars, and food commodities. This strategy, however, relies on the availability of high-quality protein sequences.

Despite the remarkable progress in sequencing technologies and the significant decrease in sequencing cost, plant genomes present unique challenges for assembly and annotation, including large variations in genome size, widely distributed repetitive sequences, high levels of ploidy and heterozygosity, and the presence of large gene families across species (25). To date, there are a limited number of protein sequences for tree nuts available in public databases [e.g., UniProt Knowledgebase (UniProt KB) and National Center for Biotechnology Information (NCBI); Table 2]. In contrast to the sequenced *Juglans regia* (English walnut) proteome with 45 745 protein sequences in UniProt KB and 56 229 protein sequences in NCBI, each of the other listed tree nut species is represented by less than 1000 available protein sequences in either database. To maximize the number of identified peptide spectra in MS/MS analyses, the implementation of proteogenomic

Table 1. Dominant allergen protein classes and tree nut allergens as recognized by the WHO/IUIS Allergen Nomenclature Subcommittee^a

Tree nut ^b	2S albumin	7S globulin	11S globulin	LTP ^c	Profilin	PR-10 ^d	Oleosin	LBP ^e
Almond			Pru du 6	Pru du 3	Pru du 4			
Brazil nut	Ber e 1		Ber e 2					
Cashew	Ana o 3	Ana o 1	Ana o 2					
Chestnut				Cas s 8 ^f		Cas s 1		
Coconut		Coc n 1						
Hazelnut	Cor a 14	Cor a 11	Cor a 9	Cor a 8	Cor a 2	Cor a 1	Cor a 12 Cor a 13	Cor a 10
Pecan	Car i 1	Car i 2	Car i 4					
Pine nut	Pin p 1	Pin k 2						
Pistachio	Pis v 1	Pis v 3	Pis v 2 Pis v 5					
Walnut, black	Jug n 1	Jug n 2	Jug n 4					
Walnut, English	Jug r 1	Jug r 2 Jug r 6 ^f	Jug r 4	Jug r 3 Jug r 8 ^f	Jug r 7 ^f	Jug r 5		

^a Additional protein allergens recognized by WHO/IUIS not reported (last updated October 2, 2018).^b Absent: macadamia nut.^c LTP = Lipid transfer protein.^d PR-10 = Pathogenesis-related protein.^e LBP = Luminal-binding protein.^f Provisional acceptance (no UniProt KB or GenBank protein listed).**Table 2. Number of tree nut protein sequences in public databases^a**

Tree nut	Species	SwissProt	TrEMBL	NCBI
Almond	<i>Prunus dulcis</i>	13	442	660
Brazil Nut	<i>Bertholletia excelsa</i>	3	92	208
Cashew	<i>Anacardium occidentale</i>	0	95	186
Chestnut	<i>Castanea sativa</i>	7	110	147
Coconut	<i>Cocos nucifera</i>	6	441	953
Hazelnut	<i>Corylus avellana</i>	7	466	631
Macadamia nut	<i>Macadamia integrifolia</i>	5	94	199
Pecan	<i>Carya illinoensis</i>	0	18	27
Pine nut	<i>Pinus strobus</i>	22	241	439
Pistachio	<i>Pistacia vera</i>	0	104	211
Black walnut	<i>Juglans nigra</i>	0	116	218
English walnut	<i>Juglans regia</i>	0	45746	56234

^a Last updated October 2, 2018.

databases that include proteins derived from translated genomic data are recommended for confident allergen identification in tree nuts.

Implementation of Comprehensive Protein Databases for Tree Nuts

In 2016, the first tree nut genome was published for English walnut (*Juglans regia*) (26). The translated reference protein sequences were first released in NCBI (BioProject PRJNA350852; 55 627 sequences) and later curated in UniProt KB (February 2018). It should be noted, at the time of data collection and analysis (September 2017), the *Juglans regia* proteome had not been added to UniProt KB; only 230 *Juglans regia* protein sequences were available in UniProt KB. To assess the effect of a comprehensive protein database on

peptide marker selection, a global proteomic analysis of raw English walnut was performed (Supplemental Information) in which MS/MS spectra were searched against two custom protein databases. The first protein database (UniProt Tree Nut) consisted of protein sequences of all major tree nut families: Anacardiaceae (1801 sequences), Arecaceae (11 863 sequences), Betulaceae (2667 sequences), Fagaceae (4264 sequences), Juglandaceae (1118 sequences), Lecythidaceae (558 sequences), Pinaceae (53 014 sequences), Proteaceae (3218 sequences), and Rosaceae (59 371 sequences) from UniProt KB. The second protein database (UniProt Tree Nut + NCBI Walnut) was built by supplementing the constructed UniProt Tree Nut database with walnut protein sequences derived from the newly sequenced and translated genome from NCBI. In total, the UniProt Tree Nut database contained 137 934 protein sequences, while the UniProt Tree Nut + NCBI Walnut database contained 193 561 protein sequences.

To alleviate the ambiguity of shared peptides and remove redundant protein identifications, proteins were inferred on the basis of parsimony and grouped into a minimal protein list (reportable proteins) sufficient to explain all identified peptides (27). For raw English walnut samples searched against the UniProt Tree Nut database, 133 protein inferences were identified, 50% of which were assigned to the genus *Prunus*, while only 7% of proteins were assigned to *Juglans regia* (Figure 2). In contrast, 277 protein inferences were identified with the UniProt Tree Nut + NCBI Walnut database with the vast majority (91%) of identifications assigned to *Juglans regia* from NCBI. The sparse number of *Juglans regia* protein sequences in UniProt KB resulted in the apparent identification of proteins primarily from *Prunus*, the only genus in the UniProt Tree Nut database with a fully sequenced genome.

A protein-centric interpretation of tree nut proteomic data can be challenging and unreliable, especially when protein sequences of target proteomes are limited. This finding is best exemplified

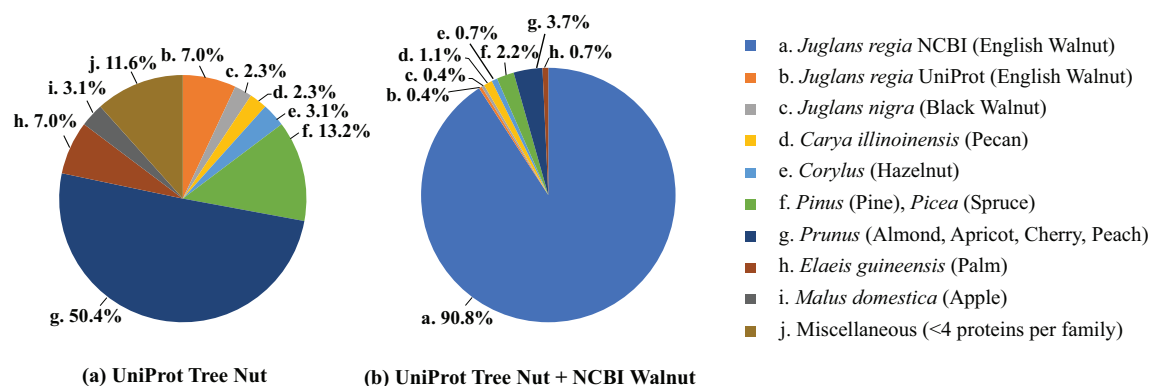


Figure 2. The percentage of parsimonious protein inferences in raw English walnut when LC-MS/MS data are searched against (a) UniProt Tree Nut or (b) UniProt Tree Nut + NCBI Walnut databases. Protein inferences required a minimum of two peptides per protein. The UniProt Tree Nut custom database includes protein sequences of all major tree nut families from UniProt KB, Anacardiaceae, Arecaceae, Betulaceae, Fagaceae, Juglandaceae, Lecythidaceae, Pinaceae, Proteaceae, and Rosaceae, as well as common protein contaminants. Translated protein sequences from the *Juglans regia* genome were downloaded from the National Institutes of Health's NCBI BioProject PRJNA350852 (55627 sequences).

in a recent study that reports the difficulty in comprehensively characterizing potential allergen contaminants in cumin (28, 29). In addition to peanut, cumin samples were screened for the presence of tree nuts; however, incomplete genome sequencing and the absence of well-characterized reference materials hindered the unambiguous discrimination between closely related *Prunus* species [i.e., *Prunus dulcis* (almond) and *Prunus persica* (peach)]. As such, the implementation of a peptide-driven protein family-based approach is recommended in the development of MS methods for tree nuts.

To date, the WHO/IUIS Allergen Nomenclature Subcommittee recognizes eight allergens for English walnut designated as Jug r (Table 1), including provisional acceptance for Jug r 6, Jug r 7, and Jug r 8. While each recognized English walnut allergen is primarily identified by a single protein isoform, the actual number of proteoforms resulting from alternative splicing, post-translational modifications, and paralogous genes

is much higher. Considering all available protein sequences, peptide identifications in raw English walnut were clustered into protein families according to shared peptide identifications. Identified protein families were categorized as allergen protein families if a WHO/IUIS Allergen Nomenclature Subcommittee recognized protein sequence for English walnut was identified within the protein family. Similarly, allergen-like protein families were defined if the annotation for a protein inference within the protein family includes a known structural or functional allergen class. Figure 3 compares the number of tryptic peptides identified to each allergen and allergen-like family in raw English walnut using UniProt Tree Nut and UniProt Tree Nut + NCBI Walnut databases. Not surprisingly, the availability of a sequenced genome for English walnut improved total peptide identifications (358 total peptides in UniProt Tree Nut database and 967 total peptides in UniProt Tree Nut + NCBI Walnut database). Specifically, the number of peptides identified for

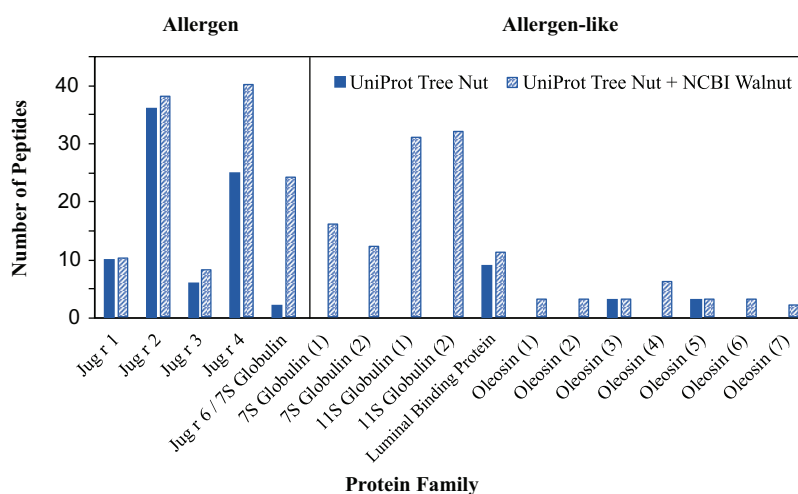


Figure 3. Number of identified tryptic peptides in raw English walnut using UniProt Tree Nut and UniProt Tree Nut + NCBI Walnut databases. Assigned peptide sequences must have a Mascot ion score greater than or equal to the identity score threshold (at $P < 0.05$) and be identified in four of the six replicate injections with a minimum of two peptides identified per protein. Indeterminate identifications, or peptides with MS/MS spectra matched to more than one amino acid sequence with equivalent scores, were manually assigned. Peptide identifications are clustered into allergen or allergen-like protein families according to shared peptide identifications. Allergen protein families are defined by identification of a WHO/IUIS Allergen Nomenclature Subcommittee recognized allergen protein sequence within the protein family; allergen-like protein families are defined by sequence annotation.

allergen and allergen-like protein families increased by an additional 41 and 110 peptides, respectively, using the UniProt Tree Nut + NCBI Walnut database. Although it is not necessary to select peptide markers representing allergenic proteins, the seed storage protein allergens (e.g., 2S albumin, 7S globulin, 11S globulin) are the most abundant proteins identified from walnut and are therefore desired targets for quantification. In addition to recognized walnut allergens, the addition of the NCBI protein sequences derived from the fully sequenced walnut genome enabled identification of unique peptide sequences in allergen-like protein families corresponding to 7S and 11S globulin seed storage proteins, a luminal-binding protein, and oleosin structural proteins, which were not identified in the UniProt-KB-only database. As access to genomic and transcriptomic data for tree nuts increases, new allergens and allergen protein

sequences will emerge. Consideration of all allergen and allergen-like protein families provides valuable supplemental evidence in the selection of target peptide markers for known and emerging allergens and for the differentiation of closely related tree nuts.

Selection of Species-Specific Peptide Markers for Tree Nuts

For allergens, candidate peptide selection is validated by empirical MS identification under defined sample preparation conditions, physiochemical peptide properties (e.g., peptide length), and specificity. MS/MS data from a global proteomic profiling of raw English walnut was searched against a custom UniProt Tree Nut + NCBI Walnut database. Table 3 and

Table 3. Total protein inferences in raw English walnut for each recognized WHO/IUIS Allergen Nomenclature Subcommittee walnut allergen protein family based on a customized UniProt Tree Nut + NCBI Walnut database^a

Protein family	Protein class	Allergen ^b	UniProt KB	NCBI walnut
Jug r 1	2S albumin	Jug r 1.0101	P93198_9ROSI ^c	XP_018824007.1 ^c
		Jug n 1.0101	Q7Y1C2_JUGNI ^c	
		Car i 1.0101	Q84XA9_CARIL ^c	
Jug r 2	7S globulin	Jug r 2.0101	Q9SEW4_9ROSI ^c	XP_018812170.1 XP_018812171.1 ^c
		Jug n 2.0101	Q7Y1C1_JUGNI	
		Car i 2.0101	B3STU4_CARIL ^c B3STU7_CARIL ^c Q15K67_CARIL	
Jug r 3	LTP ^d	Jug r 3.0101	C5H617_9ROSI ^c	XP_018814383.1 ^c XP_018814384.1 ^c XP_018814382.1 ^c
Jug r 4	11S globulin	Jug r 4.0101	Q2TPW5_9ROSI ^c	XP_018827329.1 ^c XP_018827328.1 ^c
		Jug n 4.0101	A0A1L6K371_JUGNI	
		Car i 4.0101	B5KVH4_CARIL ^c B5KVH5_CARIL	
		Cor a 9.0101	Q8W1C2_CORAV ^c A0A0A0P7E3_CORAV ^c	
Jug r 5	PR-10 ^e	Jug r 5.0101	ND ^f	ND
Jug r 6/ 7S globulin	7S globulin	Jug r 6.0101		XP_018814692.1 ^{c,g}
		Cor a 11.0101	Q8S4P9_CORAV	
Jug r 7	Profilin	Jug r 7.0101	No entry	No entry
Jug r 8	LTP ^d	Jug r 8.0101		ND
		Jug r 8.0201		ND

^a The total number of proteins within a family represents the maximum number of protein sequences, including all possible isoforms and redundant sequences, which are related by at least one shared peptide. The minimum number of proteins to explain all observable peptides is reported for each allergen protein family on the basis of parsimony (reportable proteins). Allergen protein families are defined by identification of a WHO/IUIS Allergen Nomenclature Subcommittee recognized allergen protein sequence within the protein family. Protein inferences required a minimum of two peptides per protein.

^b WHO/IUIS Allergen Nomenclature (www.allergen.org).

^c Reportable proteins.

^d LTP = Lipid transfer protein.

^e PR-10 = Pathogenesis-related protein.

^f ND = Not detected.

^g Provisional acceptance.

Supplemental Table S1 summarizes the identified allergen and allergen-like protein families, protein class, recognized WHO/IUIS Allergen Nomenclature Subcommittee allergen, and corresponding protein inferences from UniProt KB and NCBI. The total number of proteins reported within each protein family includes redundant sequences and possible isoforms, which are related by at least one shared peptide (27). Because of a high degree of sequence conservation among closely related proteins within plant families, protein sequences from other tree nuts (i.e., black walnut, pecan, and hazelnut) were clustered with English walnut allergen protein families. Highlighting the English walnut 7S globulin protein allergen Jug r 2, this protein family is characterized by four reportable proteins (seven total protein inferences) from English walnut (Q9SEW4_9ROSI, XP_018812170.1, and XP_018812171.1), black walnut (Q7Y1C1_JUGNI), and pecan (B3STU4_CARIL, B3STU7_CARIL, and Q15K67_CARIL). Among all members, Q9SEW4_9ROSI, Q7Y1C1_JUGNI, and B3STU4_CARIL are recognized protein sequences of known allergens Jug r 2.0101, Jug n 2.0101, and Car i 2.0101, respectively. To ensure the selection of marker peptides reflects an entire protein family, consideration to all available protein sequences at the order or family taxonomic level is recommended in the assembly of custom protein databases.

Analytical specificity is a critical performance parameter to evaluate possible false positives. One major drawback of

immunochemical assays for tree nut allergen detection is the high propensity of cross-reactivity with homologous proteins from other tree nuts and plant foods (30). Despite the high selectivity of MS-based assays, some current methods for the selection of species-specific peptide markers for tree nuts are limited because of the high sequence homology for closely related protein allergens combined with incomplete plant protein databases. For complex plant proteomes such as tree nuts, approaches to select allergen peptide markers are recommended to include an empirical MS evaluation of closely related allergen ingredients screened against comprehensive protein sequence databases. Currently, no commercial or certified reference materials are available for tree nut allergen analysis.

Raw and roasted varieties for all common tree nuts were analyzed by nontargeted LC–MS/MS under defined sample preparation conditions, and the resultant data were searched against the English walnut proteome (UniProt Tree Nut + NCBI Walnut) for enhanced data interrogation. Of the samples investigated, peptide identifications from English walnut (raw and roasted), black walnut (raw), pecan (raw and roasted), and hazelnut (raw and roasted) were determined to contain peptides found in recognized walnut allergen protein families. Table 4 and Supplemental Table S2 summarize the number of peptides identified for each recognized Jug r allergen protein family and allergen-like protein family. The majority of these protein families included identified peptides that were also detected

Table 4. Number of tryptic peptides identified to each recognized WHO/IUIS Allergen Nomenclature Subcommittee walnut allergen protein family in raw and roasted English walnut, black walnut (raw only), pecan, and hazelnut reference ingredients^{a,b}

Protein family	Protein class	Allergen ^c	English walnut raw/roast	Black walnut raw	Pecan raw/roast	Hazelnut raw/roast
Jug r 1	2S albumin	Jug r 1.0101 Jug n 1.0101 Car i 1.0101	13	12	7	ND ^d
Jug r 2	7S globulin	Jug r 2.0101 Jug n 2.0101 Car i 2.0101	41	45	45	4
Jug r 3	LTP ^e	Jug r 3.0101	8	2	3	ND
Jug r 4	11S globulin	Jug r 4.0101 Jug n 4.0101 Car i 4.0101 Cor a 9.0101	53	48	22	31
Jug r 5	PR-10 ^f	Jug r 5.0101	ND	ND	ND	ND
Jug r 6/ 7S globulin	7S globulin	Jug r 6.0101 Cor a 11.0101	25	15	10	23
Jug r 7	Profilin	Jug r 7.0101	ND	ND	ND	ND
Jug r 8	LTP ^e	Jug r 8.0101 Jug r 8.0201	ND	ND	ND	ND

^a MS/MS data searched against a customized UniProt Tree Nut + NCBI Walnut database.

^b Assigned peptide sequences must have a Mascot ion score greater than or equal to the identity score threshold (at $P < 0.05$) and be identified in four of the six replicate injections with a minimum of two peptides identified per protein for a single tree nut flour commodity. Allergen protein families are defined by identification of a WHO/IUIS Allergen Nomenclature Subcommittee recognized allergen protein sequence within the protein family.

^c WHO/IUIS Allergen Nomenclature (www.allergen.org).

^d ND = Not detected.

^e LTP = Lipid transfer protein.

^f PR-10 = Pathogenesis-related protein.

in pecan and/or hazelnut reference ingredients, restricting the selection of these peptides as walnut-specific peptide markers. To ensure the specificity of peptide markers, an empirical assessment of peptide uniqueness is required.

The English walnut Jug r 2 protein family was used as an example to illustrate the primary selection of species-specific peptide markers. A total of 77 tryptic peptides were reproducibly identified to members of the Jug r 2 protein family across all analyzed walnut, pecan, and hazelnut tree nut samples. Candidate Jug r 2 peptide markers were filtered by amino acid sequence length (7–20 amino acids), tryptic missed cleavage, the presence of known or identified variable modifications, and reproducible identification (4 of 6 replicate injections) in all walnut sample lysates (raw/roasted English walnut and raw black walnut). The remaining 23 peptides are summarized in Figure 4 and were further evaluated based on protein assignment and corresponding peptide identification in reference ingredients of raw and roasted pecan and hazelnut. Further removal of nonspecific peptides (Figure 4, in gray font) should be considered if the peptide is not unique to walnut protein sequence entries (e.g., QQQQCQSR) or not unique to walnut reference ingredients (e.g., EFQQCQER). Intriguingly, EFQQCQER was identified in all raw and roasted pecan and hazelnut samples without corresponding protein sequence entries in pecan or hazelnut databases, further illustrating the drawback of incomplete publicly available protein databases. Similar results are found in other walnut allergen and allergen-like protein families (data not shown).

Using English walnut as a case study, this work cautions the selection of candidate peptide markers in plant proteomes and recommends consideration of comprehensive protein sequence databases and evaluation of closely related reference food ingredients. Beyond empirical identification of peptide markers (primary peptide markers), nontargeted proteomics-based methodologies should be used to further evaluate peptides in raw and thermally processed tree nuts, cultivars, and commercial food commodities (secondary peptide markers). As MS-based methods transition toward use as confirmatory methods for allergen detection, complete protein sequence databases and well-characterized reference materials will ultimately aid in the selection of species-specific allergen peptide markers and robust analytical method development.

Moving Forward

The development of reliable methods for the detection of allergens in foods is essential to enforce food-labeling regulations and minimize the risk to allergic consumers. Although many studies have reported candidate peptide markers for allergen detection, the selection of peptide markers for tree nuts presents a number of unique challenges, including complex molecular heterogeneity (e.g., protein isoforms), closely related species, and limited availability of complete and accurately annotated protein sequence databases. In this work, we demonstrated the utility of implementing a comprehensive protein database including the sequenced genome for English walnut in combination with an empirical MS evaluation of closely related food ingredients in the selection of candidate peptide markers for allergen recognition. As tree nut genomic and transcriptomic data become available (31–34), future work will explore supplementing existing protein sequences

Peptide Sequence ^a	Protein ^b	Raw/Roast	
		Pecan	Hazelnut
ATLTLSQETR			
ATVVVVVVEGTGR		X	
CQIQEQSPER			
EEIEEIFESQMESYFVPTER			
EFQQCQER		X	X
ELSFNMMPR		X	
EQEEEEESTGR			
FFDQQEQR		X	
GAMMVPHYNSK		X	
HESEGEVK			
HNPYYFHSQSIR			
LLGFGINGENNQR			
LLQPVNNPGQFR		X	
QDPQQQYHR		X	
QFQEQR		X	
QLQEMDVLVNYAEIK		X	
QQQCQSR			
QQYEQR			
SPDQSYLR			
SSGGPISLK		X	
VFSNDILVAALNTPR		X	
VVILDANPNTFMLPHHK			
YEQCQQCER		X	X

^aAssigned peptide sequences must have a Mascot ion score greater than or equal to the identity score threshold (at $p < 0.05$) and be identified in four of the six replicate injections with a minimum of two peptides identified per protein for a single tree nut flour commodity.

^bUniProt KB or NCBI protein library entry name

Q9SEW4_9ROSI XP_018812170.1 XP_018812171.1 Q7Y1C1_JUGNI
 B3STU4_CARIL B3STU7_CARIL Q15K67_CARIL

Figure 4. Tryptic peptides identified to the 7S globulin Jug r 2 protein allergen family in all walnut sample lysates and corresponding peptide identification in reference ingredients of raw and roasted pecan and hazelnut. MS/MS data searched against a customized UniProt Tree Nut + NCBI Walnut database. Reported peptides filtered by amino acid sequence length (7–20 amino acids), tryptic missed cleavage, the presence of known or identified variable modification, and reproducible identification (four of six replicate injections). Peptides in gray font indicate nonspecific peptide markers.

with translated genomic and transcriptomic data to build more robust databases for tree nut allergen analysis. Factors that should be taken into consideration include database size, sequence fragments and redundancy, and annotation quality for the generation of such databases. We anticipate that improved protein databases will provide a more comprehensive view of tree nut protein and peptide identifications and help establish highly specific confirmatory workflows to support allergen management in the food industry.

References

- (1) Acker, W.W., Plasek, J.M., Blumenthal, K.G., Lai, K.H., Topaz, M., Seger, D.L., Goss, F.R., Slight, S.P., Bates, D.W., & Zhou, L. (2017) *J. Allergy Clin. Immunol.* **140**, 1587–1591. doi:10.1016/j.jaci.2017.04.006
- (2) Sicherer, S.H., & Sampson, H.A. (2018) *J. Allergy Clin. Immunol.* **141**, 41–58. doi:10.1016/j.jaci.2017.11.003

- (3) Gupta, R.S., Springston, E.E., Warriar, M.R., Smith, B., Kumar, R., Pongracic, J., & Holl, J.L. (2011) *Pediatrics* **128**, e9–e17. doi:10.1542/peds.2011-0204
- (4) NIAID-Sponsored Expert Panel, Boyce, J.A., Assa'ad, A., Burks, A.W., Jones, S.M., Sampson, H.A., Wood, R.A., Plaut, M., Cooper, S.F., Fenton, M.J., Arshad, S.H., Bahna, S.L., Beck, L.A., Byrd-Bredbenner, C., Camargo, C.A., Jr., Eichenfield, L., Furuta, G.T., Hanifin, J.M., Jones, C., Kraft, M., Levy, B.D., Lieberman, P., Luccioli, S., McCall, K.M., Schneider, L.C., Simon, R.A., Simons, F.E., Teach, S.J., Yawn, B.P., & Schwaninger, J.M. (2010) *J. Allergy Clin. Immunol.* **126**, S1–S8. doi:10.1016/j.jaci.2010.10.007
- (5) Clark, A.T., & Ewan, P.W. (2005) *Pediatr. Allergy Immunol.* **16**, 507–511. doi:10.1111/j.1399-3038.2005.00310.x
- (6) Croote, D., & Quake, S.R. (2016) *NPJ Syst. Biol. Appl.* **2**, 16022. doi:10.1038/npsba.2016.22
- (7) Khuda, S., Slate, A., Pereira, M., Al-Taher, F., Jackson, L., Diaz-Amigo, C., Bigley, E.C., 3rd, Whitaker, T., & Williams, K.M. (2012) *J. Agric. Food Chem.* **60**, 4195–4203. doi:10.1021/jf3001839
- (8) Parker, C.H., Khuda, S.E., Pereira, M., Ross, M.M., Fu, T.J., Fan, X., Wu, Y., Williams, K.M., DeVries, J., Pulvermacher, B., Bedford, B., Zhang, X., & Jackson, L.S. (2015) *J. Agric. Food Chem.* **63**, 10669–10680. doi:10.1021/acs.jafc.5b04287
- (9) Jayasena, S., Smits, M., Fiechter, D., de Jong, A., Nordlee, J., Baumert, J., Taylor, S.L., Pieters, R.H., & Koppelman, S.J. (2015) *J. Agric. Food Chem.* **63**, 1849–1855
- (10) Picariello, G., Mamone, G., Addeo, F., & Ferranti, P. (2011) *J. Chromatogr. A*. **1218**, 7386–7398. doi:10.1016/j.chroma.2011.06.033
- (11) Downs, M.L., & Johnson, P. (2018) *J. AOAC Int.* **101**, 146–151. doi:10.5740/jaoacint.17-0404
- (12) Sealey-Voyksner, J., Zweigenbaum, J., & Voyksner, R. (2016) *Food Chem.* **194**, 201–211. doi:10.1016/j.foodchem.2015.07.043
- (13) Korte, R., & Brockmeyer, J. (2016) *Anal. Bioanal. Chem.* **408**, 7845–7855. doi:10.1007/s00216-016-9888-y
- (14) Korte, R., Lepski, S., & Brockmeyer, J. (2016) *Anal. Bioanal. Chem.* **408**, 3059–3069. doi:10.1007/s00216-016-9384-4
- (15) Planque, M., Arnould, T., Dieu, M., Delahaut, P., Renard, P., & Gillard, N. (2017) *J. Chromatogr. A*. **1530**, 138–151. doi:10.1016/j.chroma.2017.11.039
- (16) Fisher, A., & Nawrocki, A. (2015) *Aliso* **33**, 91–110. doi:10.5642/aliso.20153302.03
- (17) Jenkins, J.A., Griffiths-Jones, S., Shewry, P.R., Breiteneder, H., & Mills, E.N. (2005) *J. Allergy Clin. Immunol.* **115**, 163–170. doi:10.1016/j.jaci.2004.10.026
- (18) Radauer, C., & Breiteneder, H. (2007) *J. Allergy Clin. Immunol.* **120**, 518–525. doi:10.1016/j.jaci.2007.07.024
- (19) Weinberger, T., & Sicherer, S. (2018) *J. Asthma Allergy* **11**, 41–51. doi:10.2147/JAA.S141636
- (20) Breiteneder, H., & Radauer, C. (2004) *J. Allergy Clin. Immunol.* **113**, 821–830. doi:10.1016/j.jaci.2004.01.779
- (21) Smeekens, J.M., Bagley, K., & Kulis, M. (2018) *Clin. Exp. Allergy* **48**, 762–772. doi:10.1111/cea.13163
- (22) Barre, A., Jacquet, G., Sordet, C., Culerrier, R., & Rouge, P. (2007) *Mol Immunol* **44**, 3243–3255. doi:10.1016/j.molimm.2007.01.023
- (23) Barre, A., Sordet, C., Culerrier, R., Rance, F., Didier, A., & Rouge, P. (2008) *Mol. Immunol.* **45**, 1231–1240. doi:10.1016/j.molimm.2007.09.014
- (24) Schein, C.H., Ivanciuc, O., & Braun, W. (2007) *Immunol. Allergy Clin. North Am.* **27**, 1–27. doi:10.1016/j.iac.2006.11.005
- (25) Claros, M.G., Bautista, R., Guerrero-Fernandez, D., Benzerki, H., Seoane, P., & Fernandez-Pozo, N. (2012) *Biology* **1**, 439–459. doi:10.3390/biology1020439
- (26) Martínez-García, P.J., Crepeau, M.W., Puiu, D., Gonzalez-Ibeas, D., Whalen, J., Stevens, K.A., Paul, R., Butterfield, T.S., Britton, M.T., Reagan, R.L., Chakraborty, S., Walawage, S.L., Vasquez-Gross, H.A., Cardeno, C., Famula, R.A., Pratt, K., Kuruganti, S., Aradhya, M.K., Leslie, C.A., Dandekar, A.M., Salzberg, S.L., Wegrzyn, J.L., Langley, C.H., & Neale, D.B. (2016) *Plant J.* **87**, 507–532. doi:10.1111/tpj.13207
- (27) Slotta, D.J., McFarland, M.A., & Markey, S.P. (2010) *Proteomics* **10**, 3035–3039. doi:10.1002/pmic.200900370
- (28) Garber, E.A., Parker, C.H., Handy, S.M., Cho, C.Y., Panda, R., Samadpour, M., Reynaud, D.H., & Ziobro, G.C. (2016) *J. Agric. Food Chem.* **64**, 1202–1211
- (29) Walker, M.J., Burns, D.T., Elliott, C.T., Gowland, M.H., & Mills, E.N. (2016) *Analyst* **141**, 24–35. doi:10.1039/c5an01457c
- (30) Cho, C.Y., Oles, C., Nowatzke, W., Oliver, K., & Garber, E.A.E. (2017) *Anal. Bioanal. Chem.* **409**, 5999–6014. doi:10.1007/s00216-017-0528-y
- (31) Jenkins, J., Wilson, B., Grimwood, J., Schmutz, J., & Grauke, L.J. (2015) *Acta. Hort.* **1070**, 101–108. doi:10.17660/ActaHortic.2015.1070.11
- (32) Mattison, C.P., Rai, R., Settlege, R.E., Hinchliffe, D.J., Madison, C., Bland, J.M., Brashear, S., Graham, C.J., Tarver, M.R., Florane, C., & Bechtel, P.J. (2017) *J. Agric. Food Chem.* **65**, 1443–1455. doi:10.1021/acs.jafc.6b04199
- (33) Moazzam Jazi, M., Seyedi, S.M., Ebrahimie, E., Ebrahimi, M., De Moro, G., & Botanga, C. (2017) *BMC Genomics* **18**, 627. doi:10.1186/s12864-017-3989-7
- (34) Cheng, L., Li, S., Chen, S., Wang, Y., Yu, M., Chen, X., Li, L., & Yin, J. (2016) *PLoS One* **11**, e0164223. doi:10.1371/journal.pone.0164223