

Human–Soybean Allergies: Elucidation of the Seed Proteome and Comprehensive Protein–Protein Interaction Prediction

Kevin Dick, Arezo Pattang, Julia Hooker, Nour Nissan, Michael Sadowski, Bradley Barnes, Le Hoa Tan, Daniel Burnside, Sadhna Phanse, Hiroyuki Aoki, Mohan Babu, Frank Dehne, Ashkan Golshani, Elroy R. Cober, James R. Green, and Bahram Samanfar*



Cite This: <https://doi.org/10.1021/acs.jproteome.1c00138>



Read Online

ACCESS |



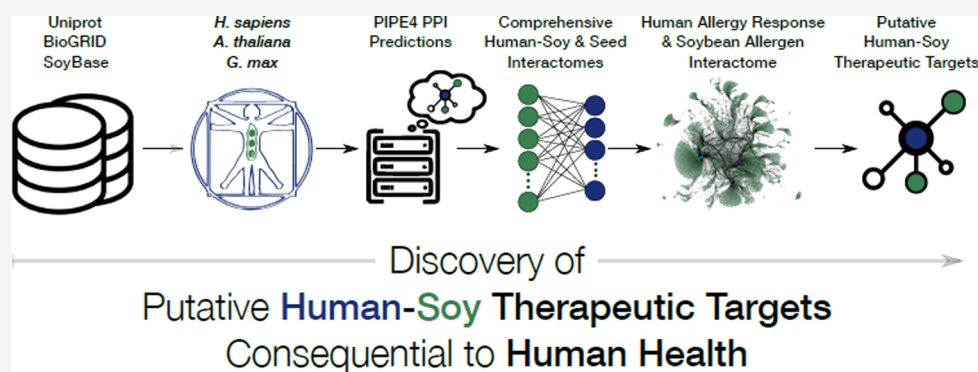
Metrics & More



Article Recommendations



Supporting Information



ABSTRACT: The soybean crop, *Glycine max* (L.) Merr., is consumed by humans, *Homo sapiens*, worldwide. While the respective bodies of literature and -omics data for each of these organisms are extensive, comparatively few studies investigate the molecular biological processes occurring between the two. We are interested in elucidating the network of protein–protein interactions (PPIs) involved in human–soybean allergies. To this end, we leverage state-of-the-art sequence-based PPI predictors amenable to predicting the enormous comprehensive interactome between human and soybean. A network-based analytical approach is proposed, leveraging *similar interaction profiles* to identify candidate allergens and proteins involved in the allergy response. Interestingly, the predicted interactome can be explored from two complementary perspectives: *which soybean proteins are predicted to interact with specific human proteins* and *which human proteins are predicted to interact with specific soybean proteins*. A total of eight proteins (six specific to the human proteome and two to the soy proteome) have been identified and supported by the literature to be involved in human health, specifically related to immunological and neurological pathways. This study, beyond generating the most comprehensive human–soybean interactome to date, elucidated a soybean seed interactome and identified several proteins putatively consequential to human health.

KEYWORDS: soybean, human allergies, seed proteome, mass spectrometry, protein–protein interactions, computational biology, PIPE4

1. INTRODUCTION

Soybean (*Glycine max* [L.] Merr.) is a high-protein oilseed used worldwide in animal feed, biodiesel, and for human consumption. While one-third of the global soybean production is provided by Brazil and the United States, with their annual exports valued at over one billion dollars, Canada ranks as the seventh-largest global producer.¹

Presently, a plethora of genome-scale -omics data are available for human and soybean including genomic sequence, transcriptomics, proteomics, and metabolomics data sets. Independently, these can offer valuable insights into the molecular processes of each organism; however, these are difficult to leverage in the context of the molecular interactions arising between the two organisms. In the context of human and soybean, elucidating the PPI network would offer insights

into human health aspects of soybean consumption. Specifically, we aim to identify soybean allergens (SAs) and the human proteins involved in the allergy response (HARs).

While food allergies (FAs) are relatively rare, they can elicit violent immune system reactions to specific food proteins. The most dangerous FA, immediate-type hypersensitivity (type I), is mediated by immunoglobulin E (IgE) antibody response and occurs in 6–8% of infants during the first 24 months.²

Received: February 17, 2021



ACS Publications

© XXXX The Authors. Published by
American Chemical Society

A

<https://doi.org/10.1021/acs.jproteome.1c00138>
J. Proteome Res. XXXX, XXX, XXX–XXX

Currently, the International Union of Immunological Societies Allergen Nomenclature Sub-Committee lists eight official SAs.³ While prior work has identified anti-soybean IgE antibodies, the allergen specificity patterns are complex, with as many as 28 known soybean proteins binding to IgE in soybean-allergic patients.^{4,5} Furthermore, as an aeroallergen, soybean seed coat proteins have been found relevant in respiratory soybean allergies resulting from the inhalation of soybean particles.⁶ Despite the study of these soybean proteins, only a rare few have been thoroughly characterized and a considerable lack of knowledge persists. Allergies manifest as the result of interactions occurring between proteins. Examining the set of all interactions (*i.e.*, the *comprehensive interactome*) promises to yield novel insights into putative SAs or candidate pharmaceutical targets within the human proteome.

Soybean can be eaten directly (*e.g.*, edamame) but is more commonly used as an ingredient in formulated foods.⁷ The soybean seed proteome encompasses proteins with multiple roles, ranging from a nutrient source during germination to a deterrent against predators. In human consumption, seed storage proteins are also the main source of allergens in soy-sensitive patients.⁸ The relationship between the physicochemical characteristics of soybean-derived foods and immune response in the form of allergens has been investigated extensively, although much is still unknown.

There are currently over 20 soybean proteins that are recognized as allergens, the majority of which are seed storage proteins including Gly m Bd 30K, Gly m Bd 60K, and Gly m Bd 28K (major SAs). Gly m Bd 60K is split into two types: glycinin (11S) and β -conglycinin (7S) accounting for roughly 80% of total allergens. β -Conglycinin has α , α' , and β subunits, while glycinin is made up of five subunits (G1–G5) that are reported to play roles as allergens.^{9,10} Gly m Bd 30K, also previously known as P34, is the main source of allergies related to soy, making up over 65% of the reactions.¹¹ This is due to Gly m Bd 30K binding oils and lipids, which contribute to its role in allergy and as a defense protein.¹² Finally, Gly m Bd 28K P28 is a minor soybean seed globulin belonging to a family of proteins responsible for a large number of legume FAs.¹³ The remainder of the seed proteome consists of protease inhibitors: Kunitz trypsin inhibitor (KTI) and Bowman–Birk trypsin inhibitor (BBI), as well as sucrose binding protein, urease, and oleosins.¹⁴ Lower abundance proteins are also present, including enzymes that mediate metabolism, which synthesize storage substances and create the structural framework of the cell.¹⁵

In addition to Gly m Bd 30K being the main source of many soy allergies, the α , α' , and β subunits of β -conglycinin have been identified as significant contributors to soybean-allergic reactions. Of the five subunits of glycinin, G1 and G2 have been identified as major allergens.¹⁶ In addition to seed storage proteins, other families have been identified as minor allergens, including KTI, profilins, oleosins, and pathogenesis-related (PR) proteins.¹⁷ Studies on different SAs utilizing comparative proteomics are beneficial to comprehending the natural variations of seed allergens.

Due to the sheer magnitude of the comprehensive interactome between *G. max* and *Homo sapiens* (~1.5 billion pairs to consider), bioinformatic analyses are promising methods to rapidly and systematically identify those proteins most likely to participate in human–soybean allergies. Those putative PPIs having a higher likelihood of interaction can thus

form the basis of testable hypotheses and warrant subsequent experimental validation.

Determining PPI networks is a key facet of molecular biology research as they provide insights into problems including the identification of gene(s) underlying disease loci,¹⁸ resolving gene function,^{19,20} and drug design.^{21,22} During the last two decades, an entire field of computational PPI prediction techniques has emerged in an effort to leverage modeling, optimization, and computing methods to complement and guide wet-lab experimentation. PPI prediction methods will typically consider a set of known PPIs (*i.e.*, previously confirmed using experimental wet-lab techniques) to optimize a model that could then be applied to any other pair of proteins to generate a score indicative of the likelihood that those proteins will physically interact.

The network-based class of predictors relies upon the connectivity of previously known PPIs to predict whether an additional connection should be incorporated between two query proteins in that network. We introduce its concepts here since a number of graph theoretical concepts are applicable to the downstream bioinformatics analyses of predicted interactomes. These predictors conventionally leverage the triadic closure principle (TCP), which originates from the observation in social network analysis wherein the more common friends held by two individuals, the more likely they are to know each other.^{23,24} For example, the STRING database leverages the TCP and consolidates evidence for protein relationships from several channels to infer PPI scores. In particular, the STRING database exemplifies how TCP applies to make inferences in both cross- and interspecies schemas.²⁵

The structure-based class of predictors generally relies on the high-resolution three-dimensional (3D) structure and domains of proteins. Generally, the richness of available physicochemical data produces high accuracy results over the set of proteins for which the data are available. The rarity of these experimentally elucidated structures, however, limits the scalability and applicability of these predictors to physicochemically diverse proteins. These limitations have previously been circumnavigated using template-based structural sequence-alignment to leverage the known 3D structures to make inferences for query proteins, which have domains that resemble known domains.²⁶ Unfortunately, it is prohibitively computationally expensive to model each putative complex with standard energy functions that minimize the conformational energy over the space of all conformation parameters.

The sequence-based class of predictors contrasts the latter methods in both requisite training data and scalability to complete interactomes. Given that they only require primary protein amino acid sequence, this class of methods can be applied proteome-wide to consider extremely large combinations of pairs of proteins. Furthermore, through computational optimizations and leveraging high-performance computing infrastructure, it became recently possible to generate the first comprehensive prediction of intraspecies interactomes (*i.e.*, the pairwise prediction of all possible PPIs) in a number of organisms.^{27,28} The protein–protein interaction prediction engine (PIPE) has been leveraged to produce comprehensive interactomes for several species including *H. sapiens*, *Mus musculus*, *Arabidopsis thaliana*, *Drosophila melanogaster*, *S. cerevisiae*, and *Caenorhabditis elegans*.^{27,29,30} These intraspecies interaction schemas have been useful for elucidating the molecular mechanisms of various organisms. The latest iteration of the PIPE algorithm, PIPE4, is also adapted to

inter- and cross-species prediction schemas with a new normalization of the original similarity-weighted score (SW score), which accounts for the species of origin of the training samples.³¹ The resulting interactome then comprises a network that can be mined for novel putative interactions.

Betweenness centrality is another concept largely employed in this research. This concept leads back to the predicted physical interactions between proteins and the network of PPIs that they form. Since the network is composed of nodes and edges, the nodes represent the proteins, while the edges represent the probable physical interactions. Betweenness centrality measures the frequency of one node appearing in all of the shortest paths between two nodes, resulting in numerical values for each node (protein) within the network and can be described as one protein's ability to monitor communication between others around it and is considered central by participating in various shortest paths.³²

The local topology of a predicted interactome can be leveraged to assign functions to unknown proteins based on their connectivity to other known proteins. In theory, a protein that interacts with several other proteins associated with a particular pathway is likely involved in that same pathway, a concept informally known as *guilt by association*. Conventionally, a protein's function can be inferred by examining the gene ontology (GO) terms associated with its interaction partner proteins. Similarly, the global topology of a predicted interactome can be leveraged to assign various measures of network centrality to a given node, capturing its relative importance with respect to other proteins.

Leveraging these concepts and the existing body of literature on human–soybean allergies and immunological pathways, a comprehensively predicted interactome can be explored from two complementary perspectives: *which soybean proteins are predicted to interact with specific human proteins* and *which human proteins are predicted to interact with specific soybean proteins*. Specifically, we can curate proteins associated with the HAR network of proteins to find soybean proteins that may interact with these and complementarily, curate proteins associated with SAs to find human proteins that may interact with these to focus our search for *de novo* PPIs consequential to human health.

Here, we leverage the state-of-the-art PPI predictor PIPE4 to perform the largest-scale investigation of the human–soy allergy response using literature-curated protein subsets of HAR and SA proteins. We experimentally elucidate the seed proteome and explore the predicted interactome to identify putative PPIs of interest using GO term analyses and determine which proteins share similar interaction profiles to ultimately form testable hypotheses for subsequent experimental validation.

2. MATERIALS AND METHODS

The last decade has seen increased computational demand, both in scale and complexity of PPI prediction tasks.^{27,33–35} The comprehensive prediction of the human–soy interactome requires evaluation of 1 533 655 878 possible PPIs, which is an estimated 10-fold increase in size over the comprehensive human interactome. Here, we use the state-of-the-art PIPE4 algorithm that was specifically developed for the application of inter- and cross-species prediction tasks. The predicted interactome is often large, and subsequent analysis can be intractable. Focusing the exploration of these interactomes can be achieved by limiting ourselves to specific subsets of proteins

of interest. Additionally, the application of statistical methods or network-based metrics can further refine the set of considered interactions. For a more intuitive exploration of the interconnectivity of these networks, an interactive variant of the static network can help researchers explore the resulting PPIs (<http://cu-bic.ca/human-soy/>).

2.1. Literature Curation of Proteins of Interest

Two subsets of known proteins involved in HAR and SA were assembled based on experimental evidence from the peer-reviewed literature. A total of 403 HAR proteins and 20 SA were extracted from the UniProt Knowledgebase (UniProtKB) (<https://www.uniprot.org/>), using the keywords “human allergy response proteins” and “soybean allergens”. Only proteins with records of the literature or curator-evaluated computational analyses were selected, and any unreviewed proteins were filtered out. Many of the soybean allergen proteins were characterized by serum IgE antibody binding studies. Of the 403 HAR proteins, a large majority fell in the immunoglobulin family. Both HAR and SA lists were then considered to identify novel putative allergens and allergy response proteins if they shared sufficiently *similar interaction profiles* (further described below) based on the generating PIPE4 predictions. These literature-curated lists are available in [Tables 1 and 2](#).

Table 1. Selected Subset of Literature-Curated Human Allergy Response Proteins. See the [Supplementary Materials](#) for the complete list of literature-curated human allergy response proteins (ST7)

human allergy response protein	Uniprot
immunoglobulin lambda constant 7	A0M8Q6
Toll-like receptor 4	O00206
coagulation factor IX	P00740
immunoglobulin heavy variable 3–13	P01766
immunoglobulin kappa constant	P01834
glucosylceramidase	P04062
cytochrome b-245 heavy chain	P04839
insulin-like growth factor I	P05019
tyrosine-protein kinase Lck	P06239
α -galactosidase A	P06280
immunoglobulin mu heavy chain	P0D0X6
lysosomal α -glucosidase	P10253
high-affinity immunoglobulin epsilon receptor subunit α	P12319
signal transducer and activator of transcription 3	P40763
high-affinity immunoglobulin epsilon receptor subunit β	Q01362
transcription factor p65	Q04206
NACH, LRR, and PYD domains containing protein 3	Q96P20
Forkhead box protein P3	Q9BZS1
nucleotide-binding oligomerization domain-containing protein 2	Q9HC29
NF-kappa-B essential modulator	Q9Y6K9

2.2. Elucidation of the Seed Proteome

Proteins were extracted from the soybean seeds with the acetone precipitation method. Briefly, 0.5 g of seeds were ground into fine powder in liquid nitrogen and homogenized with a homogenization buffer containing 20 mM Tris/HCl (pH 7.5), 250 mM sucrose, 10 mM EDTA, 1 mM protease inhibitor cocktail (Sigma-Aldrich, Germany), 1 mM PMSF, 1 mM DTT, and 1% Triton X-100 and centrifuged at 15 000g (4 °C) for 20 min. The supernatant was collected, three volumes of cold acetone was added, and then incubated at –20 °C for a

Table 2. Literature-Curated Soybean Food Allergens

soybean food allergen	Uniprot	Glycine max gene identifier	reference
GLY M IA allergen	Q9S8F3	N/A	85
Gly m 3 (Profilin-1)	O65809	Glyma05g35970.1; Glyma08g03660.1; Glyma08g03650.1; Glyma09g07800.1	86
Gly m 3 (Profilin-2)	O65810	Glyma05g35970.1; Glyma08g03660.1; Glyma08g03650.1; Glyma09g07800.1	86
Gly m 4 (Stress-induced protein SAM22)	P26987	Glyma17g03350.4; Glyma17g03350.3; Glyma17g03360.1; Glyma17g03350.2; Glyma07g37250.2	87
Gly m 5 (Glycinin G1)	P04776	Glyma19g34780.1; Glyma03g2030.1; Glyma03g32020.2	88
Gly m 5 (Glycinin G2)	P04405	Glyma03g32020.1; Glyma03g32020.2	88
Gly m 5 (Glycinin G3)	P11828	Glyma19g34780.1; Glyma03g32030.1; Glyma03g32020.1	88
Gly m 5 (Glycinin G4)	P02858	Glyma10g04280.1; Glyma13g18450.1	88
Gly m 5 (Glycinin G5)	P04347	Glyma03g32010.1; Glyma10g04280.1; Glyma13g18450.1	88
Gly m 6 (β -conglycinin- α)	POD016	Glyma10g39150.1; Glyma10g39160.1; Glyma10g39170.1; Glyma10g39170.2; Glyma20g28460.1; Glyma20g28650.1; Glyma20g28650.2;	88
Gly m 6 (β -conglycinin- β)	P25974	Glyma20g28460.1; Glyma20g28460.2; Glyma20g28640.1;	88
Gly m 7 (Seed biotin-containing protein SBP65)	Q39846	Glyma06g43640.1; Glyma12g14330.1; Glyma12g33711.1; Glyma13g36780.1	89
Gly m 8 (2S albumin)	P19594	Glyma12g34160.1; Glyma13g36400.1	90
Lectin	P05046	Glyma02g01590.1; Glyma10g01620.1	91
Gly m Bd 28 K (allergen)	Q9AVK8	N/A	92
Gly m Bd 30 K (P34) (maturing seed protein)	P22895	Glyma05g29130.1; Glyma08g12270.1; Glyma08g12280.1; Glyma08g12340.1; Glyma15g08950.1; Glyma16g17210.1;	93
Seed maturation protein PM31	Q9XET1	Glyma13g36780	94
Seed maturation protein PM30	Q9XET0	Glyma13g36780	94
Sucrose binding protein homologue S-64	Q9SP11	N/A	94
Basic 7S globulin	P13917	Glyma02g16710.1; Glyma03g30860.1; Glyma03g39940.1; Glyma06g03660.1; Glyma07g38700.1; Glyma07g38710.1; Glyma07g38720.1; Glyma10g03090.1; Glyma10g32380.1; Glyma13g27820.1; Glyma13g27820.2; Glyma13g27820.3; Glyma13g27820.4; Glyma13g27870.1; Glyma13g27870.2; Glyma13g27870.3; Glyma13g27870.4; Glyma15g11150.1; Glyma15g11160.1; Glyma15g11170.1; Glyma15g11190.1; Glyma17g01990.1; Glyma17g02000.1; Glyma17g18670.1; Glyma19g42490.1; Glyma20g35230.1; Glyma20g35240.1	94
GLY M IB allergen	Q9S8F2	N/A	85

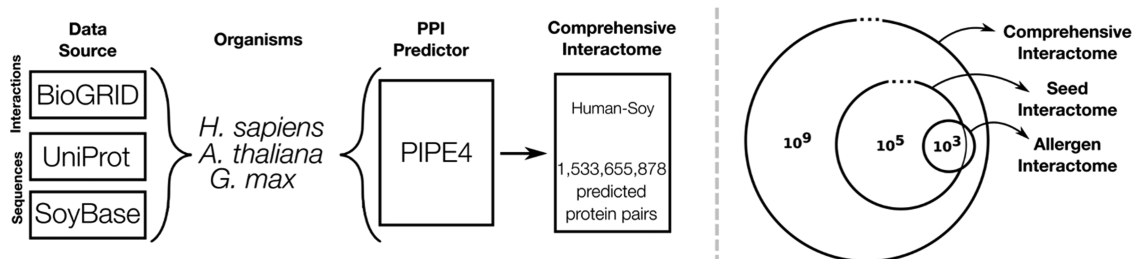


Figure 1. Prediction pipeline overview. The intraspecies training PPIs for both *A. thaliana* and *H. sapiens* were obtained from BioGrid and the corresponding protein sequences were obtained from Uniprot. The *G. max* proteome was obtained from SoyBase. The PIPE4 algorithm was used in the joint inter- and cross-species prediction schemas to predict the comprehensive interactome between all *G. max* and *H. sapiens* proteins. The putative PPIs predicted as positives were then used in subsequent network-based analyses.

minimum of 2 h. The supernatant was centrifuged, and the pellets were washed three times with cold acetone containing 1% DTT and then vacuum-dried. The protein was solubilized in a sample buffer containing 7 M urea, 2 M thiourea, 4% CHAPS, and 20 mM DTT for sodium dodecyl sulfate (SDS) electrophoresis. Two times SDS sample buffer and DTT (5 mM) were added and incubated for 30 min at 65 °C. Iodoacetamide was added for a final concentration of 10 mM and incubated at room temperature for 1 h (in the dark); it was then loaded onto a 12% sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) gel.

After staining with Coomassie Brilliant Blue, the protein gel was excised with 10 bands. The sliced gel was cut into pieces and transferred to a tube. The gels were destained with 200 μ L of methanol:50 mM NH_4HCO_3 (1:1 v/v) for 1 min with intermittent vortex mixing, and the supernatant was discarded. After dehydration for 5 min in 200 μ L of acetonitrile:50 mM NH_4HCO_3 (1:1 v/v) with intermittent vortex mixing, 200 μ L of 100% acetonitrile was added into tubes and incubated for 30 s. The samples were then dried with a Speed Vac centrifuge.

After rehydration in 20 μ L of 12 ng/ μ L trypsin in 0.01% ProteaseMAX Surfactant:50 mM NH_4HCO_3 for 10 min, the samples were overlaid with 30 μ L of 0.01% ProteaseMAX Surfactant:50 mM NH_4HCO_3 and gently mixed for several seconds and then incubated for 4 hours at 37 °C. The condensate from tube walls was collected by centrifuging at 12 000g for 10 s, mixed, and the digestion reaction with extracted peptides was transferred into a new tube. Trifluoroacetic acid (TFA) was added to a final concentration of 0.5% to inactivate trypsin. Digested samples were then desalted with a Zip-Tip C18 cartridge (Millipore). Zip-Tip was wetted with wetting solution (65% acetonitrile, 1% acetic acid in water) and then equilibrated by equilibration solution (2% acetonitrile and 1% acetic acid in water). Peptide samples were loaded by pipetting up and down 10 times with a 10 μ L pipettor in the solution. After washing using water with equilibration solution, peptide samples were eluted by 5 μ L of elution solution (65% acetonitrile, 1% acetic acid in water). Eluted peptides were air-dried and dissolved in 20 μ L of a 0.1% formic acid solution.

All samples were analyzed by nanoLC coupled to the Orbitrap Elite mass spectrometer (Thermo Fisher Scientific). Chromatographic separation of peptides was performed on a Proxeon EASY nLC 1000 System (Thermo Fisher Scientific) equipped with an in-house C18 column, 15 cm $\times$ 150 μ m ID, packed with a HxSil C18 3 μ m Resin 100 Å (Hamilton), employing a water/acetonitrile/0.1% formic acid gradient. Samples were loaded onto the column for 95 min at a flow rate of 0.30 μ L/min. Peptides were separated using 1% acetonitrile

and increasing to 3% acetonitrile in the first 2 min and then using a linear gradient from 3 to 23% of acetonitrile for 74 min, followed by a gradient from 24 to 100% of acetonitrile for 10 min and wash for 10 min at 100% of acetonitrile. The eluted peptides were directly sprayed into the mass spectrometer using positive electrospray ionization (ESI) at an ion source temperature of 250 °C and an ion spray voltage of 2.1 kV. Full-scan MS spectra (m/z 350–2000) were acquired in the Orbitrap elite at 60 000 (m/z 400) resolution. The automatic gain control settings were 1e6 for full FTMS scans and 5e4 for MS/MS scans. Fragmentation was performed with collision-induced dissociation (CID) in the linear ion trap when ion intensity was >1500 counts. The 15-most intense ions were isolated for ion trap CID with charge states ≥ 2 and sequentially isolated for fragmentation using the normalized collision energy set at 35%, activation Q at 0.250, and an activation time of 10 min. Ions selected for MS/MS were dynamically excluded for 30 s. The Orbitrap Elite mass spectrometer was operated with Thermo XCalibur software.

The MS/MS spectra of each sample were mapped to the *G. max* Wm82.a2.v1 reference + decoy protein sequences downloaded from SoyBase (<https://www.soybase.org/>)³⁶ using the SEQUEST (ver. 27, rev.9)³⁷ search algorithm for peptide and protein identification. Search parameters were set to allow for two missed cleavage sites, precursor mass tolerance ranging from -2 to 4 Da (daughter mass ion tolerance set to the default of 0), variable modification of methionine oxidation, protein N-terminal acetylation, and one fixed modification of cysteine carbamidomethylation. A stringent 1% false discovery rate threshold was used to filter candidate peptide, protein identifications. The SEQUEST search results were evaluated by the STATQUEST³⁸ probability algorithm, and only proteins with 99% probability were considered as high quality and used for further analysis. The mass spectrometry proteomics data are available via ProteomeX-change with the data set identifier "PXD023875".

2.3. Generating the Comprehensive Human–Soy Interactome

The training data used in this study are composed of experimentally validated PPIs obtained from the BioGRID (<https://thebiogrid.org/>) database. Ideally, a large number of previously validated *H. sapiens* and *G. max* interspecies PPIs would be used; however, very few of these training samples exist within the database and throughout the literature. As such, we rely upon a more complex prediction schema to leverage the existing body of research of well-studied organisms. Specifically, we use a proxy species, *A. thaliana*, given its evolutionary proximity to *G. max*. Thus, previously

validated intraspecies PPIs were obtained from BioGRID for both *H. sapiens* and *A. thaliana* and all sequences from each proteome were obtained from Uniprot (UP000005640, UP000006548); the *G. max* proteome was obtained from SoyBase (<https://www.soybase.org/>).

Adapted from the work of³¹ the PIPE4 algorithm was used to generate the comprehensive set of 766 827 939 unique predictions between *H. sapiens* and *G. max* according to the schema depicted in Figure 1. The PIPE4 algorithm is the latest version of the original PIPE algorithm introduced in ref 33 and is a uniquely optimized sequence-based PPI predictor capable of inter- and cross-species prediction schemas. Through computational optimizations and leveraging high-performance computing infrastructure, it can generate comprehensive interactomes that can comprise hundreds of millions or billions of putative interactions. An overview of the prediction pipeline used in this work is depicted in Figures 1–3. While many of the interactions from this study were originally predicted in the theoretical foundations introduced in,³¹ this work presents a quantitative exploration of those predicted interactions by considering network-based measures applied to these large-scale predictions.

2.4. Network Centrality Measures

Network-based measures of importance are used to quantify the significance of proteins within an interaction network. The concept of determining the importance of a node is consequential when evaluating the resilience of networks to perturbations. Originating from social network analysis, the *centrality* of an actor was attributed according to their position within the social structure of their environs. Since the degree of connectivity is a local quantity that may not capture the importance of a node within a holistic context, measures of centrality capture global information such as the role a node plays in the existence of paths connecting any two other nodes within the network. Here, we define the centrality measures used to interpret the predicted PPI interactomes.

Given a strongly connected and undirected graph represented as a 2-tuple, $G = (V, E)$, where V is the set of size n of all nodes and E is the set of all edges between nodes, we can define a numerical centrality measure, $c(p_i)$, of a given node, $p_i \in V$. With the degree, $\deg(p_i)$, defined as the number of edges connected to that node and the degree centrality, $c_d(p_i)$, is given as

$$c_d(p_i) = \frac{\deg(p_i)}{n - 1}$$

As the simplest measure of centrality, the degree centrality provides a quantitative measure of how interconnected a given protein is relative to other proteins in the network. A protein with a high degree centrality in a given network can be interpreted as “promiscuous” or a putative hub protein. Degree centrality alone, however, cannot inform about the surrounding interconnectivity. The closeness centrality, $c_c(p_i)$, is defined as the reciprocal of the sum of the shortest paths between p_i and p_j , $\forall j = 1, 2, \dots, n$ ³⁹

$$c_c(p_i) = \frac{n - 1}{\sum_{j \neq i} \text{dist}(p_i, p_j)}$$

where $\text{dist}(p_i, p_j)$ is the length of the shortest path between p_i and p_j , also known as a *geodesic path*. In contrast, the *eccentricity* of a given node p_i , $e(p_i)$, is the maximum path length from p_i to

any other protein and the maximum $e(p_i)$ is considered the graph diameter. Intuitively, the closeness centrality measures how *close* a given node is to all other nodes in the network. This centrality measure is useful for comparing the interconnectivity of a node relative to all others in a graph with the biological interpretation that this proximity to other nodes might suggest a similar protein function, structure, or pathway. Unfortunately, this measure of centrality becomes undefined for unconnected graphs with nodes disjoint from the giant component where distance is defined as infinite (conventionally nodes disjoint of the giant component are ignored). To correct for these undefined cases, the harmonic centrality, $c_h(p_i)$, uses the sum of the inverted distances rather than the inverted sum of distances

$$c_h(p_i) = \frac{1}{n - 1} \sum_{j \neq i} \frac{1}{\text{dist}(p_i, p_j)}, \text{ where } \frac{1}{\infty} = 0 \text{ for disjoint nodes}$$

The betweenness centrality, $c_b(p_i)$, is a more complex measure of centrality and interpreted as the amount of influence a protein has over pairwise connections between other proteins by counting the fraction of shortest paths going through a given node. The greater the betweenness centrality of a node, the greater it perturbs the network if eliminated or modified. This assumes that all connections are of equal importance along all of the shortest paths for each pair

$$c_b(p_i) = \frac{\left(2 \sum_s \sum_t^{g_{st}(p_i)}\right) / g_{st}}{(n-1)(n-2)}, \quad s \neq t \neq i$$

where $g_{st}(i)$ is the number of geodesics from a source protein, p_s , to a target protein, p_t , containing the protein p_i , and g_{st} is the total number of geodesics between p_s and p_t . While each centrality measure is useful for the interpretation of the importance of proteins within the network, the betweenness centrality has traditionally been used to interpret the influence of a given protein on the network. With the assumption that mutation or changing expression of a certain protein might result in its elimination from the network, the betweenness centrality seeks to quantify the impact of this elimination upon the remaining network. This necessitates the assumption that all PPIs share equal importance and that the network does not contain redundancies, both weak in the context of PPI networks.

2.5. Investigation of Proteins from the Human Allergy Response and Soybean Allergen Networks

The predicted interactome for the subset of human proteins contributing to the HAR was extracted, as was the predicted interactome for the subset of soybean proteins associated with SAs. Static visualizations of both networks were created and are depicted in Section 3.2, where the network edges are weighted by the PIPE4 similarity-weighted score and the node size (in reference to the network visualizations in Section 3.2) representing the betweenness centrality. The network layout was achieved using a mixture of force-based algorithms and edge expansion to avoid overlapping nodes. As such, a number of variable-sized clusters form containing proteins sharing similar sets of interactions.

The static depiction of these networks is visually appealing and highlights global features of the resulting interactomes; however, such visualizations hold little value in communicating

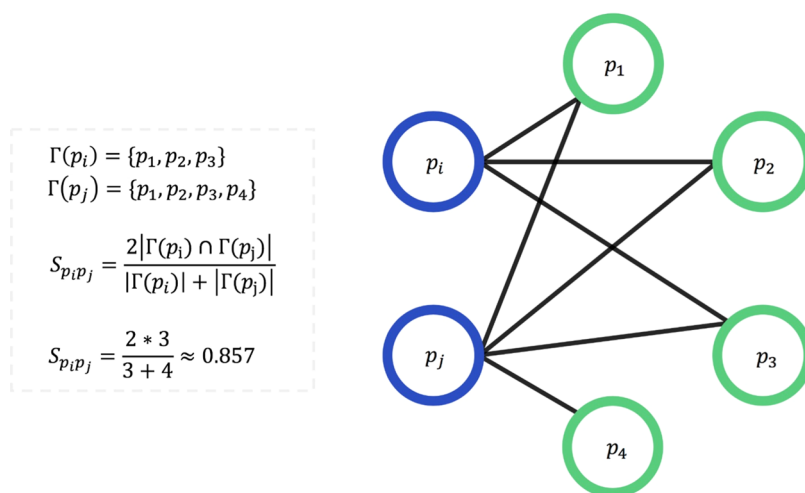


Figure 2. Example application of the Sørensen–Dice coefficient for two arbitrary human proteins. Graphical representation of leveraging the TCP to compute pairwise similarity of human proteins p_i and p_j based on their predicted interaction with soybean proteins p_1 , p_2 , p_3 , and p_4 . This network also exemplifies the small interaction set effect where the elimination of even one predicted interaction (say p_i – p_3) would result in a considerable drop in the similarity coefficient: $S'_{p_i p_j} = (2 * 2)/(2 + 4) \approx 0.667$; a drop of $\sim 20\%$.

protein-specific or local topological information. Thus, an interactive variant of these networks is made available at <https://cu-bic.ca/human-soy/> for a more enriching experience of navigating the local connectivity of individuals or groups of proteins.

These curated subsets of proteins and their predicted interactors are used in subsequent explorations of gene ontologies (GOs), as well as determining which proteins among the complete proteome share similar interaction profiles as those extracted.

2.6. Gene Ontology Term Analyses

The Protein ANALysis THrough Evolutionary Relationships (PANTHER) Classification System (version 11) was used to map and analyze the GO terms⁴⁰ for each set of proteins predicted to interact with the curated sets of proteins. For the HAR network, we examined each of the molecular functions, biological processes, and cellular components of the predicted soybean proteins and vice versa for the SA network. To each of these six resulting GO term analyses, we independently applied a statistical over-representation test using a binomial statistical comparison and Bonferroni correction for multiple testing. The over-representation test compares the count of proteins in the query list against the reference proteome. Since we are testing a large quantity of GO terms for significant enrichment among our query proteins, multiple testing corrections are required. To identify the GO terms that are significantly over-represented among our predicted subset, we sorted our results by positive fold difference. The results of these analyses are available in ST1–ST6.

2.7. Proteins with Similar Interaction Profiles

From the curated set of known SAs and proteins involved in the HAR, we sought to identify other proteins in the respective proteomes having similar interaction profiles. We define a protein's interaction profile as the set of putative interactions involving that protein. Here, the motivating hypothesis originates from the TCP rooted in the study of social networks: the observation that the greater the number of common friends of two individuals, the higher the likelihood that they know each other. This principle has been adapted to biological interactomes on the assumption that proteins

sharing many similar known interactors are likely to interact themselves. Recent work has suggested that this direct interpretation may not necessarily hold for biological networks and the authors propose a *path length of three* (L3) alternatives for network-based PPI prediction.⁴¹ Interestingly, the evolutionary and network structural basis for the proposed L3 method is the same that motivates the similarity-weighted score of the PIPE4 algorithm used in this work.

To clarify, in this network, we leverage TCP not to identify putative PPIs but rather to identify new SA or HAR proteins by identifying those proteins within one proteome that have similar interaction profiles as known SA or HAR proteins. We measure this similarity using the Sørensen–Dice similarity coefficient and apply it in a one-to-all examination of the similarity profiles between each of the curated proteins and the remainder of the proteome. These values are then sorted into rank order, and the results are plotted using one-to-all score curves, a visualization adapted from the reciprocal perspective (RP) framework.^{42,43} Each of these curves is used to provide qualitative evidence for the meaningfulness of two proteins having a high degree of similarity based on the comprehensively predicted interactome. The exploration of the interaction network for those previously known allergens enables the identification of novel putative allergens.

We define the interactor function, $\Gamma(\bullet)$, which takes as input a protein, p_i , and returns a set of size γ containing all proteins participating in a predicted interaction with p_i . For example

$$\Gamma(p_i) = \begin{cases} \emptyset, & \text{if } \gamma = 0 \\ \{p_j\}, & \text{if } \gamma = 1 \\ \{p_j, \dots, p_\gamma\}, & \text{if } \gamma > 1 \end{cases}$$

To identify proteins sharing similar interaction profiles to those identified here, we used the Sørensen–Dice coefficient and computed all pairwise values for each set of interactors for protein p_i with the interactors for each protein in the remainder of the proteome $p_j \forall j$

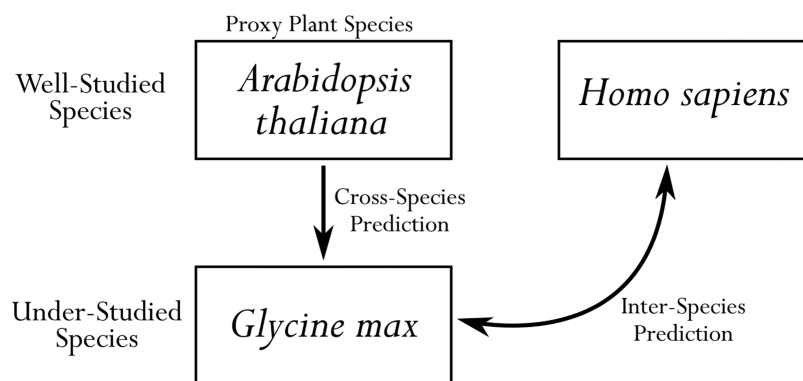


Figure 3. Prediction schema overview. Given the lack of experimentally validated PPIs between human and soybean, we leverage an evolutionarily related model organism *A. thaliana* for a joint inter- and cross-species prediction schema.

$$S_{p_i p_j} = \frac{2|\Gamma(p_i) \cap \Gamma(p_j)|}{|\Gamma(p_i)| + |\Gamma(p_j)|}$$

where $S_{p_i p_j} \in [0,1]$ and a value of 1 indicates identical interaction profiles, while 0 indicates completely disjoint interaction profiles. We are interested in those proteins that share many of the same interacting partners as those predicted to participate in each of the HAR and SA networks while having a few predicted interactions outside of these networks. This shared coefficient of similarity exhibits a trade-off between a given protein's shared interactors and a protein's completely predicted interactome: proteins will have a lower similarity coefficient if they share interactors, with *many* of the interactions in the HAR and SA networks; however, they also have many predicted interactions outside of these networks; conversely, proteins may also have a lower similarity coefficient if they only share a few interactions of the HAR and SA networks; however, they do not have predicted interactions outside of these networks.

This analysis can be biased for those proteins that interact with a small number of highly connected (hub) proteins since a large number of proteins in the proteome are likely to also interact with these few hub proteins. For this reason, it is imperative that we consider not only the top-ranking proteins but also the distribution of similar profiles across the entire proteome. Conversely, a *small interaction set effect* may be observed wherein, for small interaction set sizes (generally <100), the rank-order distribution of all $S_{p_i p_j}$ for a given protein p_i might exhibit a stepwise descent resulting from the attribution of large changes in the similarity coefficient to the single prediction or rejection of a single protein from the HAR or SA networks (Figure 2).

2.8. Validation Experiment Predicting Interactions between Related Legumes and Human Allergen Targets

To validate that the PIPE4 algorithm and proposed schema will generally assign a higher-predicted score between known allergens within the seed proteins of related legumes and human target proteins, a complementary computational positive control experiment was run. We first collected the Uniprot protein accession numbers and amino acid sequences (with at least 20 amino acids in length) for all known allergens from related legumes with homologues to soybean proteins ($m = 39$), as well as their known human target proteins ($n = 21$). Leveraging a similar prediction schema, intraspecies *A. thaliana* training PPIs were considered as a proxy for the collection of

legume proteins and combined with intraspecies *H. sapiens* training PPIs to generate predictions for all legume–human pairs ($m \times n = 819$). The training PPIs were extracted from BioGRID following the same protocol outlined in Section 2.3; we tabulated all legume proteins considered in this analysis in Supplementary Table 10 (ST10), which includes the Uniprot accession number, protein name, legume organism name, and the length of the protein sequence in a number of amino acids.

The PIPE4 algorithm was used to generate the all-to-all predictions between the $m = 39$ legume proteins and all proteins within the human proteome (which contains the subset of 21 target proteins of interest). As a comparison of predicted scores between the 39 legume proteins and the 21 human target proteins, we randomly sampled 21 human proteins (distinct from the 21 human proteins of interest) from the complete human proteome for each of the legumes. This set of randomly sampled human proteins is assumed to score lower against the 39 legumes than the 21 human target proteins, with the assumption that our PIPE4 schema will assign a greater score to those allergens and allergy response pairs. We visualized the distribution of these comparative predictions and quantified our experiment by reporting the number of instances where, for each legume protein, the maximum predicted value of the target human protein exceeds that of the maximum predicted value for the randomly sampled proteins.

3. RESULTS

The comprehensive prediction of the human–soybean interactome required a complex inter- and cross-species prediction schema given the limited availability of experimentally validated interspecies PPIs between *H. sapiens* and *G. max*, as well as intraspecies *G. max* interactions (Figure 3). The evolutionarily similar and well-studied *A. thaliana* was therefore leveraged as a proxy organism for *G. max*, and the interactome was predicted by computing the PIPE4 scores for all 1 533 655 878 protein pairs. However, due to the symmetrical relationship of an undirected PPI, only half of these PPIs are required to generate the comprehensive interactome (e.g., $AB \equiv BA$, where A and B are proteins).

3.1. Literature-Curated Proteins of Interest: Human Allergy Response and Soybean Allergens

Literature-curated proteins of interest involved in both HAR and SA were identified to find candidate proteins consequential to human health. The literature-curated HAR proteins have been extracted from studies associating the human proteins

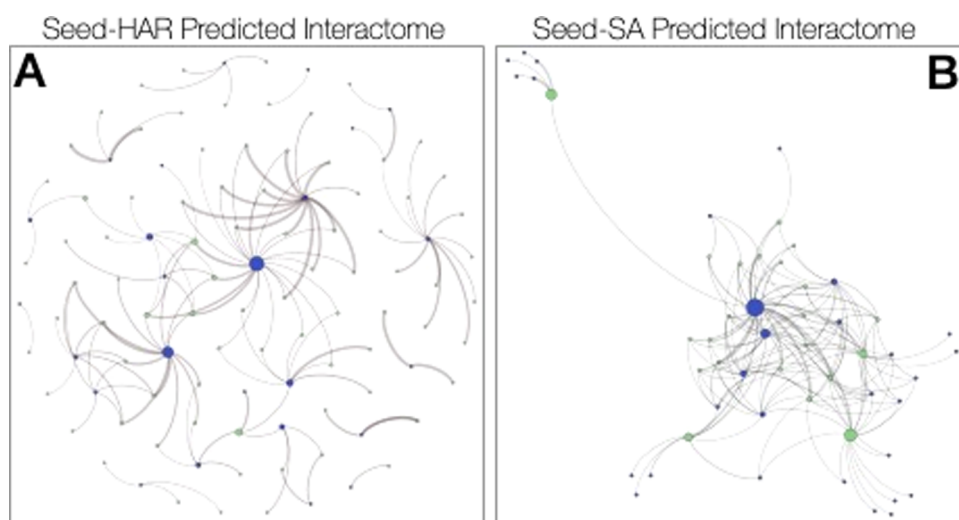


Figure 4. Intersected HAR- and SA-predicted interactomes with the predicted seed interactome. Node size is proportional to the node betweenness centrality, and edge weight is proportional to the predicted PPI score. In panel (A), the seed–HAR-predicted network of 4091 proteins and 19 351 edges comprises 11.17% of soybean proteins and 88.83% of human. The graph diameter (ignoring disjoint components) is 7. In panel (B), the seed–SA-predicted network of 107 proteins and 134 edges comprises 76.64% of soybean proteins and 23.36% of human. The graph diameter is 11. The complete list of predicted results for both interactomes is included in the Supplementary Material (SF1 and SF2). The interactive and queryable version of this network is available at <http://cu-bic.ca/human-soy/>.

with the general allergic response. The SA proteins were curated from the literature based on known SAs, allergens from other species related to soybean, and soybean proteins that are homologous to known allergens from a variety of related species. Contrastingly, between the literature-curated proteins involved in HAR and SA, there are significantly more proteins identified as being involved in the HAR ($n = 403$) than identified SAs ($n = 20$). A select subset of literature-curated HAR can be found in Table 1, with the complete list available in Supplementary Table 7 (ST7), and all identified SAs are found in Table 2.

It is apparent that some protein families were of greater significance and occurred frequently. Among the HAR proteins (ST7), the immunoglobulin family was found to have the highest frequency of occurrence, including both heavy (P0D0Y5, A0A0C4DH62, P01766, P01763, *etc.*), kappa (P01834, P0D0X7, A0A0A0MT89, P0602, *etc.*), and lambda (A0A075B6K0, A0A075B6J6, A0A075B6K6, A0A087WSX0, P01721, *etc.*) variables. Other protein families found to have a relatively high frequency of occurrence among the HAR included T-cell β chain (A2NZL7, A2NZL9, A2NZM1, A2NZL8, *etc.*) and glucosylceramidase (P04062, P10253, D3DV87, B7Z6S9, *etc.*). Among the curated SA (Table 2), the majority were those previously identified as having a significant contribution to allergens including the α , α' , and β subunit of β -conglycinin, including the cupin family: Glyma10G39150.1, Glyma10G39160.1, Glyma10G39170.1, Glyma20G28460.1, Glyma19G34780.1, Glyma03G32020.2, Glyma19G34780.1, Glyma03G32010.1, *etc.* To a lesser extent, SAs such as profilin (Glyma18G32740.1 (uncharacterized), Glyma05G35970.1 (profilin-2), Glyma05G35960.1 (profilin), *etc.*), seed biotinylated proteins (Glyma06G43640.1 (seed biotinylated protein), Glyma12G14330.1 (late embryogenesis abundant domain-containing protein), Glyma12G33710.1 (late embryogenesis abundant domain-containing protein), *etc.*), albumin (Glyma12G34160.1 (protease inhibitor/seed storage/LTP family) and Glyma13G36400.1 (protease inhibitor/seed storage/LTP family), and lectin (Glyma02G01590.1

(lectin protein kinase family protein) and Glyma10G01620.1 (legume lectin domain) were also identified.

3.2. Extracting the Predicted Subnetworks

From the comprehensively predicted interactome (high-scoring protein pairs were extracted using a cutoff determined at a 99.95% specificity), those interactions predicted to interact with one of the HAR or SA proteins were extracted and depicted as an interconnected graph in Supplementary Figure 1 (SF1; the interactive and queryable version of this network is available at <http://cu-bic.ca/human-soy/human>) and Supplementary Figure 2 (SF2; the interactive and queryable version of this network is available at <http://cu-bic.ca/human-soy/soy>). Each protein is represented as a node, and the edges between nodes represent the predicted PPIs. Node size represents its betweenness centrality (where a larger node has a larger betweenness centrality), and edge weight is similarly determined by the predicted likelihood of interaction. The predicted HAR network has a graph diameter of 10 and comprises 5168 proteins (97.5% soybean; 2.5% human) and 8054 edges (SF1). The predicted SA network has a graph diameter of 6 and comprises 154 proteins (49.4% soybean; 50.6% human) and 438 edges (SF2). While seemingly dense and difficult to interpret as static figures, their interactive and queryable variants can be much more intuitive to explore relationships among neighboring nodes. These interactive variants are available at <http://cu-bic.ca/human-soy/> (Figure 4).

To further refine the search for putative interactions within the seed proteome, soy proteins predicted to interact with HAR that were also among the seed proteome were identified ($n = 134$), with the top-30 predictions listed in Table 3. Similarly, the human proteins strongly predicted to interact with SA and that fall among the seed proteome were also identified ($n = 151$), with the top-30 predictions listed in Table 4. A complete list of these interactions is available in the SF1 and SF2; their network representations are depicted below. The predicted interactions among the soybean seed proteome ($n = 328$) and the intersection of the soybean seed proteome

Table 3. Top-30 Predictions from the Human Allergen Response Proteome ($n = 403$, $m = 8054$). Italics (human) denote proteins of the source proteome^a

rank	human	soybean	PIPE4 SW Score
1	<i>O95678</i>	Glyma02g15610.1	37.86
2	<i>Q15555</i>	Glyma04g33890.2	36.65
3	<i>Q15555</i>	Glyma06g20550.1	36.61
4	<i>Q15555</i>	Glyma17g10720.1	36.23
5	<i>P02538</i>	Glyma02g15610.1	35.73
6	<i>P48668</i>	Glyma02g15610.1	35.22
7	<i>P04259</i>	Glyma02g15610.1	35.13
8	<i>Q15555</i>	Glyma17g10720.2	34.97
9	<i>P13647</i>	Glyma02g15610.1	31.97
10	<i>Q15555</i>	Glyma04g33890.1	31.88
11	<i>Q15555</i>	Glyma11g05560.1	31.27
12	<i>Q15555</i>	Glyma01g39720.1	31.18
13	<i>P51671</i>	Glyma04g02190.1	30.45
14	<i>Q05513</i>	Glyma13g20380.1	30.27
15	<i>P15918</i>	Glyma15g19300.1	30.22
16	<i>Q99616</i>	Glyma04g02190.1	26.39
17	<i>P07737</i>	Glyma06g03840.1	22.95
18	<i>Q04206</i>	Glyma06g00200.2	21.32
19	<i>O95678</i>	Glyma03g33450.1	21.19
20	<i>P01308</i>	Glyma17g31680.1	20.68
21	<i>P02538</i>	Glyma03g33450.1	20.67
22	<i>P48668</i>	Glyma03g33450.1	20.44
23	<i>P04259</i>	Glyma03g33450.1	20.28
24	<i>O14543</i>	Glyma13g20380.1	20.17
25	<i>P06239</i>	Glyma14g36620.2	20.15
26	<i>Q04206</i>	Glyma06g02150.1	19.79
27	<i>P06733</i>	Glyma06g36810.1	19.77
28	<i>Q04206</i>	Glyma12g09190.1	19.52
29	<i>Q04206</i>	Glyma14g25210.1	19.38
30	<i>Q04206</i>	Glyma04g36100.1	19.35

^aThe value n denotes the number of unique proteins of the source proteome, while the value m denotes the number of unique pairs in the predicted interactome.

and HAR interactome were also identified ($n = 107$) (available in the [ST7](#) and [ST8](#)), and the top-30 predictions can be found in [Tables 5](#) and [6](#).

For each network, the top-30 proteins, sorted by decreasing betweenness centrality, are tabulated in [Tables 7](#) and [8](#). All centrality values are normalized to the numeric range [0,1]. Each network comprises a mix of human and soybean proteins; however, we are most interested in the proteins predicted to interact with the literature-curated subsets of proteins such as soybean proteins that interact with the HAR proteins and human proteins that interact with the SA proteins. To narrow our attention to those predicted proteins of the complement proteome, we emphasize the top-10 proteins by betweenness centrality with respect to the organism. From the HAR network, the top-10 soybean proteins are listed in [Table 9](#); from the SA network, the top-10 human proteins are ranked among the top 30 in [Table 8](#) (see the emboldened protein names), with key candidates supported by literature-founded immunological techniques ([ST9](#)).

Two proteins from the *G. max* proteome have a large betweenness centrality ranked among the top 30 in the predicted HAR networks, Glyma18g48660.1 (protein of unknown function) and Glyma11g05670.2 (cyclin-dependent kinase inhibitor 1) ([Table 7](#)). Expanding this analysis to a less

Table 4. Top-30 Predictions from the Soybean Allergen Proteome ($n = 76$, $m = 438$). Italics (soybean) denote proteins of the source proteome^a

rank	human	soybean	PIPE4 SW score
1	P0S067	<i>Glyma13g36400.1</i>	9.74
2	P0S067	<i>Glyma20g28650.2</i>	8.98
3	P0S067	<i>Glyma20g28660.1</i>	8.77
4	P0S067	<i>Glyma20g28650.1</i>	8.77
5	P0S067	<i>Glyma12g14330.1</i>	8.50
6	P0S067	<i>Glyma17g03350.2</i>	8.41
7	P0S067	<i>Glyma06g43640.1</i>	8.15
8	P0S067	<i>Glyma17g03350.1</i>	8.15
9	P0S067	<i>Glyma03g32010.1</i>	8.07
10	P0S067	<i>Glyma17g03350.4</i>	8.02
11	P0S067	<i>Glyma17g03350.3</i>	8.02
12	P0S067	<i>Glyma10g39150.1</i>	8.01
13	P0S067	<i>Glyma17g03360.1</i>	7.79
14	P0S067	<i>Glyma07g37240.1</i>	7.66
15	P0S067	<i>Glyma13g36780.1</i>	7.66
16	P0S067	<i>Glyma10g39170.1</i>	7.63
17	P01040	<i>Glyma16g17210.1</i>	7.63
18	P0S067	<i>Glyma13g18450.1</i>	7.60
19	P0S067	<i>Glyma12g34160.1</i>	7.57
20	P0S067	<i>Glyma07g37240.3</i>	7.51
21	P0S067	<i>Glyma03g30860.1</i>	7.46
22	P01040	<i>Glyma05g29130.1</i>	7.22
23	P0S067	<i>Glyma15g11170.1</i>	7.20
24	P0S067	<i>Glyma10g39170.2</i>	7.19
25	P01040	<i>Glyma15g08840.1</i>	7.15
26	P0S067	<i>Glyma02g01590.1</i>	7.11
27	P0S067	<i>Glyma03g32020.2</i>	7.02
28	Q0VD86	<i>Glyma10g09890.1</i>	7.01
29	P0S067	<i>Glyma15g11190.1</i>	6.95
30	P0S067	<i>Glyma20g28640.1</i>	6.89

^aThe value n denotes the number of unique proteins of the source proteome, while the value m denotes the number of unique pairs in the predicted interactome. The top-30 predictions here are also the top 30 among the seed–SA interactome, and, thus, the seed–SA table of results is not duplicated. See [SF3](#) for the complete seed–SA list of interactions.

conservative subset, the top-10 unique (nonisomorphic) soybean proteins, sorted by betweenness centrality, are listed in [Table 9](#). Interestingly, among the top-30 proteins of the SA network, the top-10 human proteins are contained ([Table 8](#)), including one KTi 3 (P49223), three amyloid-related (P0S067, Q06481, and P51693), and two mothers against decapentaplegic (MAD) homologues (P84022 and Q15796), which are known to participate in the regulation of the 14-3-3 family of proteins, corroborating findings discussed later.

In addition to the holistic interpretation of each unique protein in the context of all others, we can expand the analysis to investigate the functions these predicted sets of proteins belong to, relative to the reference proteome.

3.3. Gene Ontology Term Analyses

We sought to examine the associated protein functions for the proteins predicted to interact with the human and soybean subsets of interest. The Protein ANalysis THrough Evolutionary Relationships (PANTHER) GO database⁴⁰ was used to analyze the 2973 mappable *G. max* proteins predicted to interact with the subset of HAR proteins and the 72 mappable *H. sapiens* proteins predicted to interact with the SA proteins.

Table 5. Top-30 Predictions from the Soybean Seed Proteome ($n = 328$, $m = 19\,352$). Italics (soybean) denote proteins of the source proteome^a

rank	human	soybean	PIPE4 SW score
1	O14531	<i>Glyma15g06100.1</i>	51.10
2	Q14195	<i>Glyma15g06100.1</i>	48.48
3	Q14194	<i>Glyma15g06100.1</i>	48.31
4	Q16555	<i>Glyma15g06100.1</i>	47.40
5	Q14117	<i>Glyma15g06100.1</i>	46.79
6	O60361	<i>Glyma07g13710.1</i>	45.11
7	Q9BPU6	<i>Glyma15g06100.1</i>	43.56
8	Q6FHN3	<i>Glyma07g13710.1</i>	43.06
9	P22392	<i>Glyma07g13710.1</i>	43.06
10	Q13232	<i>Glyma07g13710.1</i>	43.03
11	P15531	<i>Glyma07g13710.1</i>	42.96
12	Q14376	<i>Glyma05g38120.6</i>	40.14
13	Q14376	<i>Glyma05g38120.5</i>	40.14
14	P21580	<i>Glyma04g10820.2</i>	39.54
15	Q14376	<i>Glyma11g01940.2</i>	39.49
16	Q9UQ84	<i>Glyma04g10820.2</i>	38.69
17	P04075	<i>Glyma02g38730.1</i>	38.17
18	P04406	<i>Glyma06g18110.6</i>	38.02
19	P04075	<i>Glyma10g07710.1</i>	37.76
20	P05062	<i>Glyma02g38730.1</i>	37.49
21	P04075	<i>Glyma14g36850.1</i>	37.28
22	P04406	<i>Glyma04g36870.2</i>	37.27
23	P04406	<i>Glyma06g18110.5</i>	37.25
24	P04406	<i>Glyma06g18110.3</i>	37.24
25	P04406	<i>Glyma18g01330.2</i>	37.11
26	P05062	<i>Glyma10g07710.1</i>	37.10
27	Q9UK53	<i>Glyma04g10820.2</i>	36.95
28	P04406	<i>Glyma18g01330.1</i>	36.95
29	P04406	<i>Glyma11g37360.1</i>	36.88
30	P04406	<i>Glyma06g18110.4</i>	36.83

^aThe value n denotes the number of unique proteins of the source proteome, while the value m denotes the number of unique pairs in the predicted interactome (ST8).

Ontologies were examined for each available category: molecular function (MF), biological process (BP), cellular component (CC), and pathways. The top-level ontology for each category is depicted in Figure 5, and a description and complete tabulation of the statistical over-representation tests used are available in ST1–ST6.

3.4. Similar Interaction Profiles

In this work, we present the predicted interactome for a subset of proteins curated for their importance and relevance to the human–soy allergy response. In the previous sections, we examined the two predicted interactomes to identify proteins with various measures of connectivity (e.g., betweenness centrality) as putative targets for subsequent therapeutic design and development. We then holistically examined the members of this predicted interactome using their associated GO terms. Having established that the members of this predicted interactome are generally relevant to the human–soy allergy response, we now leverage the comprehensive interactome to identify candidate proteins that share similar interaction profiles.

Expanding our exploration to consider the complete predicted interactome, we leveraged the Sørensen–Dice (SD) similarity coefficient, enabling the pairwise comparison of proteins based on their interaction profile. We define a

Table 6. Top-30 Predictions from the Intersection of the Soybean Seed Proteome–Human Allergy Response Interactome ($n = 107$, $m = 134$)^a

rank	human	soybean	PIPE4 SW score
1	P02788	<i>Glyma10g28890.2</i>	16.43
2	P02788	<i>Glyma10g28890.1</i>	15.85
3	Q99650	<i>Glyma19g22780.1</i>	14.88
4	Q99650	<i>Glyma18g01330.2</i>	14.68
5	Q99650	<i>Glyma11g37360.1</i>	14.65
6	Q99650	<i>Glyma18g01330.1</i>	14.59
7	Q99650	<i>Glyma06g18110.3</i>	14.49
8	Q99650	<i>Glyma04g36870.2</i>	14.47
9	P04179	<i>Glyma10g39050.1</i>	14.37
10	Q99650	<i>Glyma06g18110.1</i>	14.37
11	Q99650	<i>Glyma04g36870.1</i>	14.36
12	Q99650	<i>Glyma06g18110.4</i>	14.18
13	P04179	<i>Glyma11g00890.1</i>	14.13
14	Q99650	<i>Glyma06g18110.6</i>	14.10
15	Q99650	<i>Glyma06g18110.5</i>	14.069
16	P07737	<i>Glyma19g22850.1</i>	13.019
17	Q05513	<i>Glyma04g10820.2</i>	12.90
18	P19878	<i>Glyma05g26400.1</i>	12.85
19	P19878	<i>Glyma08g09310.1</i>	12.63
20	P40763	<i>Glyma17g10710.1</i>	12.39
21	Q99650	<i>Glyma06g18110.2</i>	12.34
22	P07737	<i>Glyma10g35700.1</i>	11.95
23	P07737	<i>Glyma05g24110.1</i>	11.68
24	P07737	<i>Glyma19g07240.2</i>	11.63
25	P07737	<i>Glyma19g07240.1</i>	11.63
26	Q05513	<i>Glyma04g10820.3</i>	11.51
27	P15918	<i>Glyma10g29000.1</i>	11.42
28	Q05513	<i>Glyma04g10820.1</i>	11.34
29	Q05513	<i>Glyma04g10820.4</i>	11.26
30	Q04206	<i>Glyma09g34760.3</i>	10.69

^aThe value n denotes the number of unique proteins in the interactome, while the value m denotes the number of pairs (ST8).

protein’s *interaction profile* as the set of all proteins it is predicted to interact with. We are interested in identifying those proteins that might share *similar interaction profiles* with a given protein of interest. From the predicted HAR and SA interactomes, we can expand our analysis beyond this isolating context to query all proteins within each interactome for similar interaction profiles based on the comprehensive interactome. This comparison can be performed for all pairs of proteins, and those sharing highly similar interaction profiles (many shared interactions with the intersection, very few interactions in the symmetric difference) would score a value close to one, while those with highly dissimilar interaction profiles would score close to zero (zero itself if the interaction profiles are disjoint).

Computing the SD coefficient between one protein and all others in the proteome and then plotting the rank-order distribution of these values allow the appraisal of the “uniqueness” of a given interaction profile throughout the interactome. A small number of proteins that share a high degree of similarity, where all other proteins in the proteome are relatively low or zero, suggest a specific set of shared interactions indicative of a common function or pathway. These specific types of interaction profiles potentially warrant additional investigation to elucidate pathways of interest. Contrarily, a protein exhibiting a high degree of similarity with

Table 7. Network Measures of Protein Connectivity for the Top 30 by Betweenness Centrality from the Human Allergy Response Network with all Soybean Proteins^a

protein id	org.	$\varepsilon(p_i)$	$\deg(p_i)$	$c_b(p_i)$	$c_b(p_i)$	protein name
Q04206	H	5	1312	0.501	0.393	Transcription factor p65
P40763	H	5	824	0.436	0.237	Signal transducer and activator of transcription 3
P06239	H	5	632	0.412	0.188	Tyrosine-protein kinase Lck
Glyma18g48660.1	S	6	55	0.454	0.133	Protein of unknown function
P07737	H	6	327	0.36	0.093	Profilin-1
Q9Y6K9	H	5	322	0.369	0.084	NF-kappa-B essential modulator
P15918	H	7	330	0.361	0.065	V(D)J recombination-activating protein 1 (RAG-1)
Q05513	H	7	332	0.365	0.063	Protein kinase C zeta type (nPKC-zeta)
P56537	H	7	205	0.342	0.061	Eukaryotic translation initiation factor 6 (eIF-6)
O14543	H	5	252	0.355	0.051	Suppressor of cytokine signaling 3 (SOCS-3)
P10599	H	5	145	0.34	0.051	Thioredoxin (Trx)
P06733	H	5	207	0.349	0.048	α -enolase
A0A0A0MT76	H	7	121	0.306	0.042	Immunoglobulin lambda joining 1
Q8N138	H	7	122	0.329	0.035	ORM1-like protein 3
P02788	H	7	138	0.311	0.035	Lactotransferrin
Q15555	H	5	155	0.344	0.033	Microtubule-associated protein RP/EB family member 2
P00451	H	7	48	0.308	0.03	Coagulation factor VIII
O00626	H	7	95	0.277	0.028	C–C motif chemokine 22
P02768	H	7	74	0.319	0.025	Serum albumin
P04179	H	7	82	0.327	0.024	Superoxide dismutase
P00709	H	9	53	0.201	0.020	α -lactalbumin
P51809	H	7	59	0.251	0.019	Vesicle-associated membrane protein 7 (VAMP-7)
Glyma11g05670.2	S	6	31	0.403	0.019	Cyclin-dependent kinase inhibitor 1
P26038	H	5	98	0.333	0.018	Moesin
P51671	H	5	61	0.319	0.016	Eotaxin
P05019	H	7	61	0.313	0.015	Insulin-like growth factor I (IGF-I)
O00329	H	7	54	0.292	0.015	Phosphatidylinositol 4,5-bisphosphate 3-kinase catalytic subunit delta isoform
P01308	H	5	69	0.323	0.014	Insulin
P24394	H	7	42	0.272	0.014	Interleukin-4 receptor subunit
P19878	H	7	69	0.325	0.014	Neutrophil cytosol factor 2

^aCentrality measures are defined in the [Materials and Methods](#) section. Emboldened (soybean) are the predicted proteins of the complement proteome.

a large proportion of the proteome suggests that there is a core set of predicted interactions that are shared among the majority of proteins and are thereby not unique. The resulting distribution of these one-to-all rank-order curves can also exhibit a characteristic knee, which can help suggest a protein-specific similarity cutoff; the rare few highly similar proteins should be considered, while those among the low similarity baseline are not. We employ TCP as an exploratory tool rather than a PPI predictor to quantify the similarity of *function* and/or *pathway* rather than *interaction*.

The one-to-all curves (and their corresponding violin plots) were computed for each of the predicted proteins of the HAR and SA networks. Nine proteins were selected from each set to depict the diversity of their one-to-all distributions ([Figure 6](#)).

Among the nine soybean proteins, a number of general one-to-all distributions are observed and qualitatively categorized. Both Glyma02g16710 (aspartyl protease) and Glyma13g36780 (late embryogenesis abundant domain-containing protein/LEA domain-containing protein) share a *linearly decreasing* phenotype suggestive of a large cluster of interactions common to the majority of proteins. Similarly, Glyma10g32380 (aspartyl protease) depicts a *stepwise decreased* phenotype indicative of a small common cluster of interactions. A *long-tailed* distribution is exemplified with Glyma05g35970 (profilin) where a small proportion of PPI share a high degree of similarity and a large proportion of proteins share a low,

nonzero, degree of similarity. Three proteins, Glyma08g03650 (profilin), Glyma08g12340 (cysteine proteinase cathepsin F), and Glyma16g17210 (cysteine proteinase cathepsin F), each depict a *piecewise linear* phenotype, a more extreme form of the long-tailed distribution, wherein definitive points in the curve delineate a sharp junction in the degree of similarity. Glyma08g03650 (profilin) depicts two such points (~ 0.80 , ~ 0.20), and both Glyma08g12340 (cysteine proteinase cathepsin F) and Glyma16g17210 (cysteine proteinase cathepsin F) depict one (~ 0.25 and ~ 0.30 , respectively). Finally, the L-shaped phenotype is the most extreme form of the latter two distributions where only a few rare proteins share any above-zero similarity and the remainder of the proteome shares none at all.

Among the predicted human proteins within the SA network, similar distributions are observed ([Figure 6](#)). The top row depicts rarity, where Q96P20 (NACHT, LRR, and PYD domains containing protein 3) exemplifies the majority of cases with an L-shaped distribution, P06315 (immunoglobulin kappa variable 5-2) exemplifies a piecewise linear distribution due to a consistently decreasing density of similarities, and P13647 (keratin, type II cytoskeletal 5) with a unique patch of similarity centered around ~ 0.75 . The middle row depicts three proteins with some mixture of stepwise (O95394, Q8TDQ1), piecewise (Q15555), and long-tailed (all three) distributions. Finally, the bottom row depicts three proteins

Table 8. Network Measures of Protein Connectivity for the Top 30 by Betweenness Centrality from the Soybean Allergen Network with all Soybean Proteins^a

protein id	org.	$\varepsilon(p_i)$	$\deg(p_i)$	$c_h(p_i)$	$c_b(p_i)$	protein name
P05067	H	4	71	0.705	0.559	Amyloid- β A4 protein
Glyma05g35970.1	S	3	17	0.489	0.199	Profilin
Q06481	H	4	50	0.581	0.124	Amyloid-likeprotein 2
Glyma08g03650.1	S	3	9	0.454	0.102	Profilin
A8MXD5	H	4	7	0.379	0.092	Glutaredoxin domain-containing cysteine-rich protein 1
Q17RA5	H	4	7	0.379	0.092	Putative uncharacterized protein C21orf62-AS1
Glyma13g27820.1	S	5	8	0.430	0.077	Aspartyl proteases
Glyma20g28650.2	S	5	17	0.471	0.061	α subunit of β -conglycinin
Glyma17g02000.1	S	5	7	0.426	0.054	Aspartyl protease
Glyma12g34160.1	S	5	9	0.435	0.043	Napin-type 2S albumin 1 precursor
P84022	H	4	31	0.474	0.042	Mothers against decapentaplegic homologue 3 (MAD homologue 3)
P63279	H	4	16	0.430	0.036	SUMO-conjugating enzyme UBC9
Glyma06g43641.1	S	5	13	0.454	0.033	Late embryogenesis abundant (LEA)-related
Glyma13g18450.1	S	5	5	0.417	0.028	Glycinin gy5
Q15796	H	4	25	0.446	0.027	Mothers against decapentaplegic homologue 2 (MAD homologue 2)
Glyma15g11170.1	S	5	10	0.439	0.026	Aspartyl protease
Glyma05g35970.1	S	5	7	0.359	0.025	profilin
Glyma08g12340.1	S	5	7	0.426	0.025	P34 probable thiol protease-like
Glyma08g12270.1	S	5	6	0.422	0.025	34 kDa maturing seed vacuolar thiol protease precursor
P51693	H	6	30	0.464	0.024	Amyloid-like protein 1
Glyma16g17210.1	S	5	6	0.422	0.020	Cysteine proteinase cathepsin F
Glyma12g33711.1	S	5	10	0.441	0.019	Late embryogenesis abundant protein
Q502X0	H	4	3	0.336	0.019	MORN repeat-containing protein 2
Glyma07g37240.3	S	5	8	0.430	0.018	Stress-induced protein SAM22
P49223	H	6	25	0.442	0.017	Kunitz-type protease inhibitor 3
Glyma12g14330.1	S	5	7	0.428	0.017	Late embryogenesis abundant (LEA)-related
Glyma03g32020.2	S	5	8	0.432	0.017	Cupin
Glyma20g35230.1	S	5	4	0.413	0.014	Exocyst complex component (Section 5)

^aCentrality measures are defined in the [Materials and Methods](#) section. Emboldened (human) are the predicted proteins of the complement proteome. The top-10 unique human proteins are captured among this top 30 in the network.

Table 9. Top-10 Unique Soybean Proteins Predicted to Interact with the Human Allergy Response Network by Betweenness Centrality^a

protein id	UniRef100	$\deg(p_i)$	$c_h(p_i)$	$c_b(p_i)$	protein name
Glyma18g48660.1	N/A	55	0.454	0.133	Protein of unknown function
Glyma11g05670.(1,2,3,4)	Q6T2Z3	31	0.403	0.019	Cyclin-dependent kinase inhibitor
Glyma03g03070.1	G7KQ56	2	0.236	0.012	T-complex protein 1 subunit theta
Glyma10g26400.1	B9HW82	2	0.318	0.012	Heavy-metal-associated domain-containing family protein
Glyma16g21430.1	Q84ZV0	14	0.331	0.011	R 14 protein
Glyma10g25770.1	G7K0X2	3	0.319	0.010	Mitotic checkpoint protein BUB3
Glyma01g39580.1	N/A	21	0.397	0.010	Ubiquitin-protein ligase
Glyma05g17900.1	Q0WMT6	21	0.397	0.010	ARM-repeat/tetratricopeptide repeat-like protein
Glyma17g18570.1	N/A	21	0.397	0.010	Ubiquitin-protein ligase
Glyma15g01130.1	N/A	3	0.324	0.009	Protein of unknown function

^aDuplicate isoforms are merged into a single entry, and the centrality measures of the highest are reported and the isoform number is emboldened.

with the extreme form of the L-shaped distribution, however, each with a small, nonzero similarity shared across thousands of proteins in the proteome: P01308 (insulin), P51671 (eotaxin), and P56537 (eukaryotic translation initiation factor 6).

In addition to the one-to-all rank-order distributions, violin plots are used to depict the density of similarities across the [0,1] range. These are useful for examining specific portions of the one-to-all curves, which have over- or under-represented degrees of similarity, particularly with respect to both network cluster size and number. The density of the distribution phenotypes corresponding to the all-to-one curves described

above can be compared from a multifaceted perspective (Figure 6).

3.5. Predicting Interactions between Related Legumes and Human Target Proteins

The PIPE4 algorithm was used to generate predictions between the protein sequences for $m = 39$ related legume proteins and $n = 21$ human target proteins. As in the general schema presented throughout this manuscript, *A. thaliana* was considered as a proxy organism for *G. max* and the comprehensive prediction matrix of scores was produced including the ($m \times n = 819$) scored pairs that are reported in [Supplementary Table 11 \(ST11\)](#). With the assumption that the

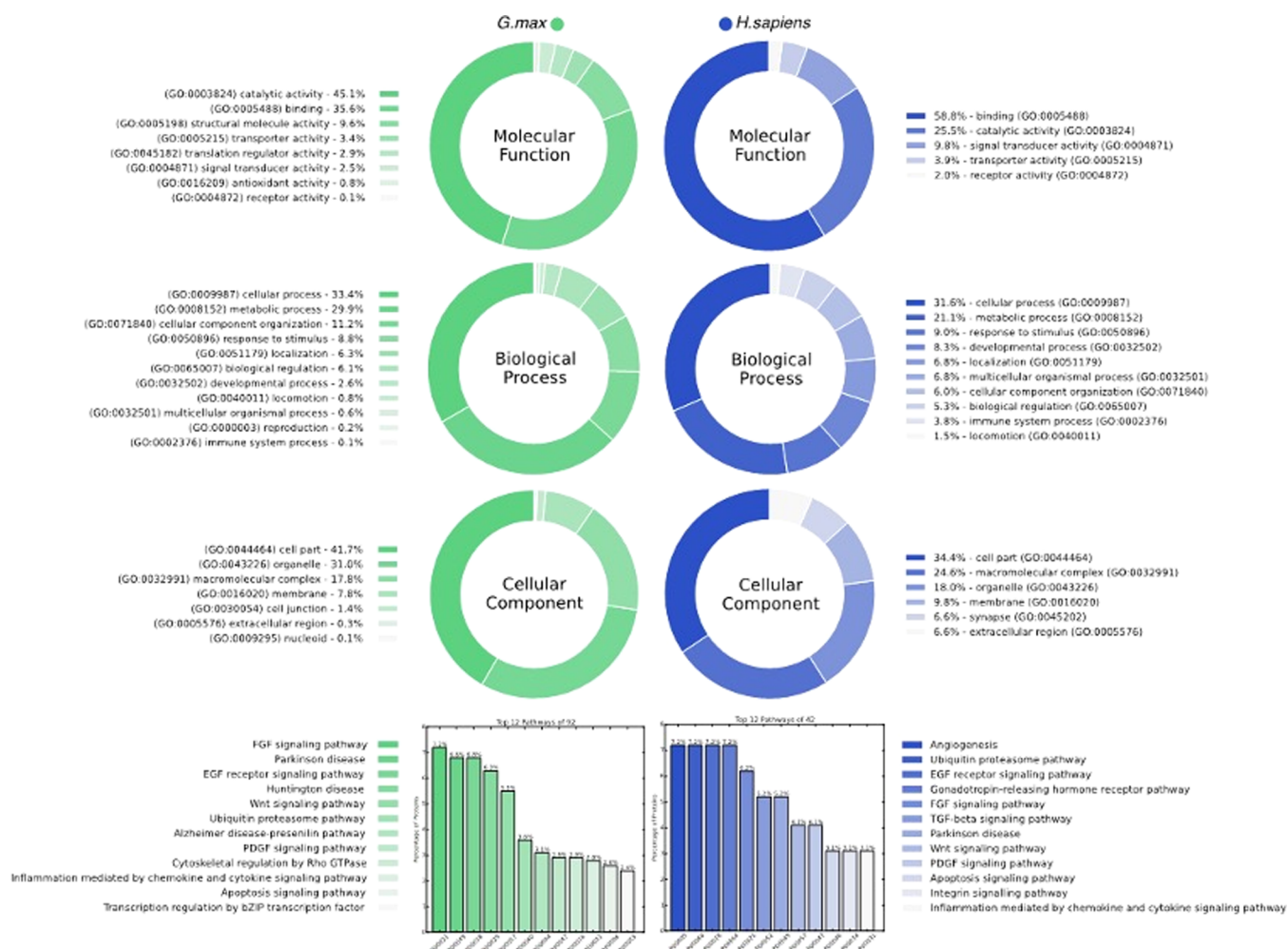


Figure 5. Summary of the PANTHER GO term analyses of the predicted interactions. Only the highest level of the ontological hierarchy is depicted here. A complete tabulation of the statistically over-represented ontologies is available in [Supplementary Tables S1–S6](#). Intrafigure colors are linearly interpolated based on GO term enrichment.

legume–human target predictions would generate higher scores for the pairs of known interactions between these allergens and allergy response proteins, we compared our predictions to a random sample of scores from the comprehensive prediction matrix. For each of the $m = 39$ legume proteins, a random sample of 21 human protein pairs (equivalent to the number of target proteins) were extracted and their similarity-weighted prediction score was compared.

To demonstrate that PIPE4 generally predicts a higher score between the $m = 39$ legume proteins and the $n = 21$ human proteins, a comparative figure of violin plots was created and is visualized in [Figure 7](#), where we observe that, on average, the scores generated among the target human proteins ([Figure 7](#) in blue) score higher against the known legume allergens as compared to the equivalently sized randomly sampled set of human proteins ([Figure 7](#) in orange). We assume that the maximum scoring prediction between a given legume allergen and the human target proteins represents a real interacting pair. Consequently, we further quantify our validation experiment by counting the number of instances for each of the 39 legume proteins where the maximum scoring target human protein score exceeds that of the randomly sampled human protein. We observe that 26/39 or approximately 66% of the legume proteins outscore their comparison, which is over half of the proteins considered. The proteins for which

this is observed include A0A1S6WN66, A5Z1Q8, B0YIU5, B1PYZ4, B2CMA3, B2ZGS2, B3EWP3, B3EWP4, B3IXL2, B5TYU1, B6CEX8, B6CG41, C0HJZ1, D3K177, E5G077, P43238, Q08169, Q0GM57, Q0PKR4, Q2YHR1, Q45W79, Q5XXQ4, Q6IWG5, Q6VT83, Q8LKN1, and Q9SQI9. We would further nuance that the comparison of the maximum scoring pairs is a broad assumption and the observation that target human proteins tend to score higher than those randomly sampled pairs further validates the methodology presented in this work.

4. DISCUSSION

Soybean is increasing as one of the most important crops worldwide due to its high nutritional values used for both human consumption and animal feed. Recent amendments to the nutritional labeling legislature require soybean allergen labeling, as soybean has been recognized as a common food source allergen and is among the “big eight” in the United States and top 14 in European Union. In the United States, soybean accounts for ~90% of IgE allergic reactions that are food-mediated.⁴⁴ At least 28 SAs have been identified through IgE binding in soybean-sensitive individuals including β -conglycinin, glycinin, Gly m Bd 28K, and Gly m Bd 30K or P34 as being the major allergen in the seed proteome.⁴⁵ Hence,

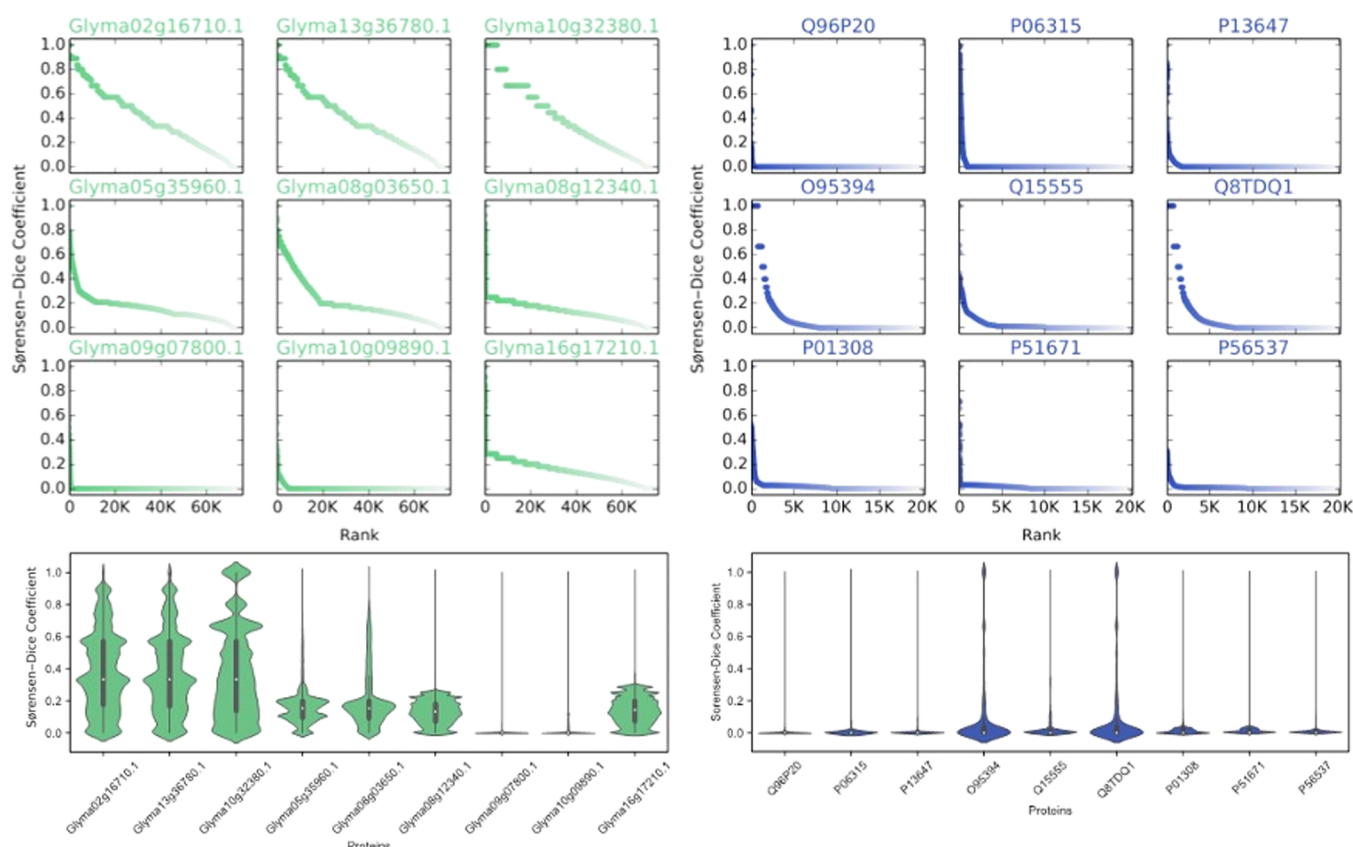


Figure 6. Example Sørensen–Dice similarity coefficient one-to-all curves and violin plots. Example of nine one-to-all curves depicting the similarity of the interaction profile of one protein against all others in the proteome for each of the predicted soybean proteins of the HAR network (left; green) and the predicted human proteins of the SA network (right; blue). Colors are linearly interpolated.

it is no surprise that a large number of the identified HAR proteins are IgE antibodies. Since FAs are potentially deadly reactions triggered by the immune system through food protein antigens, in the last two decades, substantial advancements have been made in pinpointing allergenic soy proteins.⁴⁶

Both human and soybean genomes have been widely researched separately. In this paper, a state-of-the-art sequence-based PPI predictor, PIPE4, was leveraged using *A. thaliana* as a proxy for the soybean legume to generate predictions for all possible pairs between *G. max* and *H. sapiens*. The interspecies interactome (~1.5 billion PIPE4 predictions), GO term analysis, pathway analysis, along with known data sets of HAR, SA, and soybean seed proteome, were essential in allowing us to perform a large-scale investigation into the human–soybean allergy response, identifying the top interspecies interacting partners and revealing a number of biological mechanisms consequential to human health.

One key element of our analysis is considering the *similar interaction profile* in a one-to-all framework of Sørensen–Dice scores based upon betweenness centrality, which is best explained as the number of times one node is a part of the shortest path connecting the different nodes within the network. Betweenness centrality accounts for distant protein influences, whether direct or indirect, while observing the node on the shortest path as the one with the largest effect between them all. Our focus was on the proteins with high betweenness centrality and highly similar interaction profiles with those HAR and SA proteins within the literature-curated lists.⁴⁷

4.1. Multiperspective Prevention of Allergies

A fundamental facet of interspecies predictions is the capability to observe the network from two complementary perspectives: predicting the interaction of soybean proteins with human proteins versus predicting the interaction of human proteins with soybean proteins. From one perspective, we curate a set of human proteins, which are known to be associated in some way with the HAR from prior literature: the HAR proteins. From the complementary perspective, we curated a set of soybean proteins, which are known to be associated in some way with SAs from prior literature: the SA proteins.

Producing the comprehensive set of interactions for these two organisms enables a holistic network-based analysis of putative PPIs, which may contribute to the manifestation of soybean-resultant allergenic responses in humans. A more targeted investigation of those proteins that are elucidated to exist within the subset seed proteome and which are predicted to interact with any of the HAR and SA proteins promises to reveal putative PPIs that may contribute to human–soybean allergies. Moreover, the identification of high-confidence putative PPIs (following the review of the existing literature) could form the basis of testable hypotheses and a potential target for designed therapeutics. From the human protein perspective, a small inhibitory molecule may be designed to competitively bind proteins with high propensity to the soybean interaction partner. Conversely, from the soybean perspective, genetic engineering may be leveraged to modify the site of interaction, mediating the interaction.

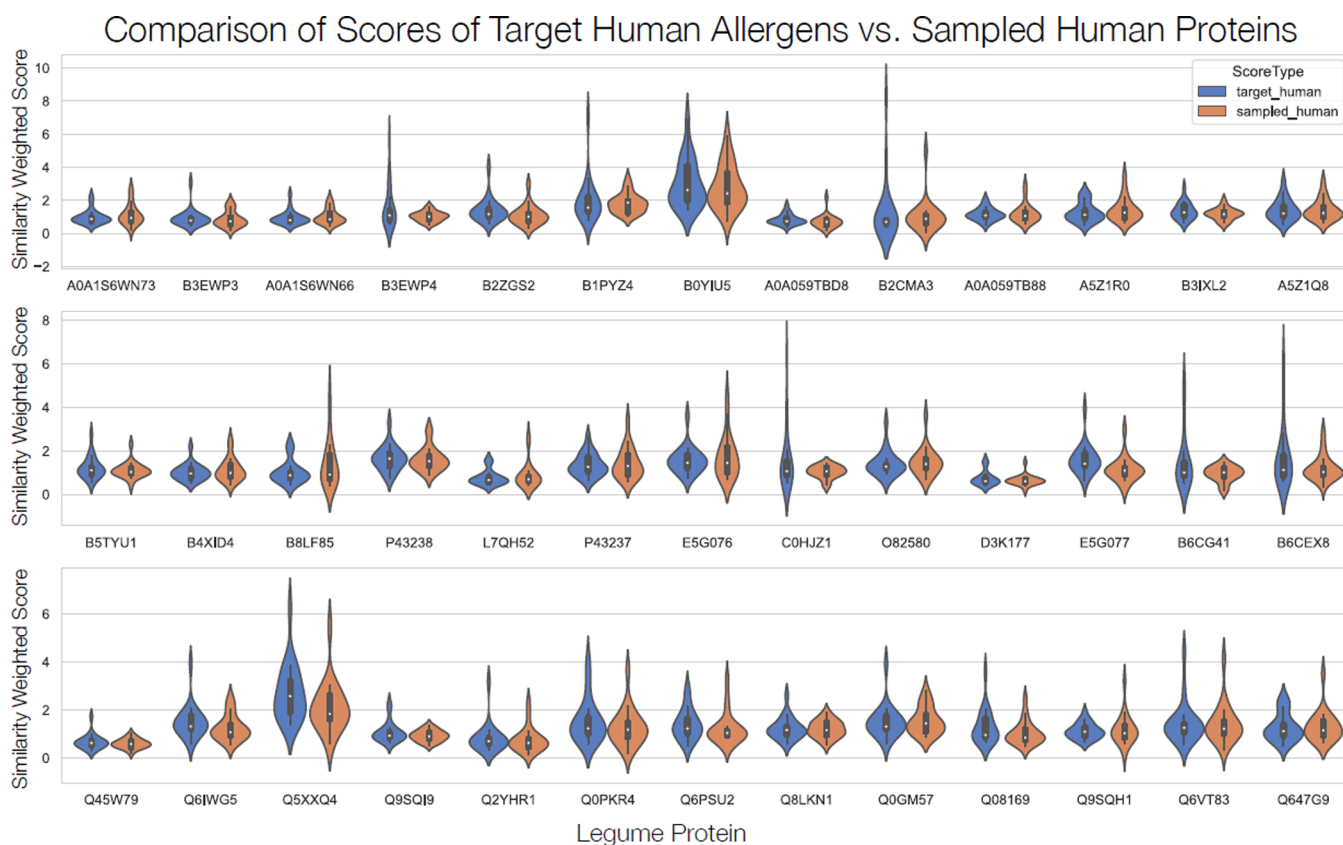


Figure 7. Validation study demonstrating that 39 legume proteins tend to score higher in similarity-weighted PIPE4 score than an equivalent random sample of human proteins. The PIPE4 algorithm was used to predict the score between each legume protein and all human proteins. The subset of 21 scores with each human target protein was extracted to make the target human set of similarity-weighted scores. An equivalent number of randomly sampled scores with human proteins were curated to comprise the sampled human set of similarity-weighted scores. Generally, we observe that the target human SW scores are greater than the randomly sampled human protein SW scores.

Similarly, we sought to identify specific proteins that may be targeted to maximally disrupt the interactome with the premise that targeted modification or elimination of these proteins could disrupt protein interactions. This was achieved using a variety of network-based measures of centrality applied to the predicted HAR and SA subnetworks.

4.2. Network Analyses Reveal Neurological Proteins with Large Betweenness Centrality

Extraction and network-based analyses of the HAR- and SA-predicted subnetworks revealed a number of predicted interactions involving proteins having high betweenness centrality. Since these networks were extracted based on the predicted interactors of the curated HAR and SA sets of proteins, we can expect those proteins to have a large betweenness centrality (e.g. the set of all proteins predicted to interact with protein p_x , $\Gamma(p_x) = \{p_1, p_2, \dots, p_n\}$ will inflate the betweenness centrality of p_x relative to each $p_y \in \Gamma(p_x)$). Therefore, we are neither interested in the human proteins of the set of HAR proteins nor the soybean proteins of the set of SA proteins. Rather, we are interested in the soybean proteins with high betweenness centrality among those predicted in the HAR network and, similarly, interested in the human proteins with high betweenness centrality among those predicted in the SA network (Tables 7–9).

Among the human proteins found to have high betweenness centralities within the SA network, there were several identified as being involved in neurological or neurodevelopmental processes. Within the SA network, the human protein amyloid-

β (P05067) is predicted to interact with 71 of the curated soybean proteins, resulting in the largest betweenness centrality among all nodes in the network, despite being a protein from the complement proteome (Table 2). An extensive body of literature has linked the accumulation of this protein in the brain to the development of Alzheimer's disease (AD).^{48–50} Furthermore, within this network, amyloid-like proteins (APLP) 1 and 2 (P51693, Q06481), the Kunitz-type protease inhibitor 3 (P49223) (of the family of Kunitz-type domains), and two mothers against decapentaplegic homologues 2 and 3 (SMAD2 and 3: Q15796 and P84022, respectively) also share high betweenness centrality and have a linkage to neurological development. Both amyloid-like proteins 1 and 2 (APLP1 and APLP2, respectively) were identified to be involved in synaptic processes important for learning and memory function. Cognitive decline has been attributed to the absence of functional APLP1 and APLP2, and the accumulation of these proteins is correlated to AD. The SMAD2 and SMAD3 proteins have been identified as mediators of the transforming growth factor β (TGF- β) signaling. The deregulation of the TGF- β family signaling pathway has been identified to be associated with various neurological disorders, including AD, Parkinson's disease (PD), Huntington's disease (HD), and amyotrophic lateral sclerosis (ALS), to name a few.⁵¹ To a lesser extent, the human protein SUMO-conjugating enzyme UBC9 (P63279) was also found to share a large betweenness centrality within the SA network but was not predicted to interact with as many curated soybean proteins. However,

much like the previously mentioned proteins, UBC9 was also identified to be involved in neurodegenerative diseases, including HD, PD, and AD.^{52,53}

Within the HAR network, two soybean proteins were found to have high betweenness centralities (Table 7), including Glyma18G48660 (uncharacterized) with its *Arabidopsis* ortholog identified as a SUMO conjugation enzyme 1 and Glyma11G05670 (cyclin-dependent kinase inhibitor 1) (Cdk1). Both have been linked to neurological or neurodevelopmental processes. SUMO stands for Small Ubiquitin-like Modifier, which are proteins that modify the function of other proteins by covalently attaching and detaching to their lysine (Lys) residues. They are vital to ensuring eukaryotic cells function normally.^{54,55} The post-transcriptional modification process is referred to as SUMOylation and is responsible for numerous processes including apoptosis, response to stress, cell growth, tumorigenesis, and transcriptional regulation. They have also been implicated in various diseases including heart disease, brain ischemia, and cancer and are critical in regulating synaptic function and neuronal integrity.^{56,57}

Previous literature has associated Cdks in cell cycle control, with Cdk1, Cdk2, Cdk4, and Cdk6 expressed during cell division and Cdk5 expressed in postmitotic cells. There is evidence suggesting a relationship between Cdks and neurodegenerative diseases, including AD and PD, acting as potential mediators of ischemic neuronal death.⁵⁸ Cdk1 specifically has been identified as being responsible for phosphorylating key substrates necessary for cytokinesis upon activation, with the literature correlating the increased expression of Cdk1 in response to ischemic-induced neuronal cell death.^{58–60}

The human proteins that have been identified as sharing high betweenness centralities within the SA network, while also predicted to interact with many curated soybean proteins, including amyloid- β , APLP1 and 2, KTi 3, and SMAD2 and 3, were also identified to share interaction with several soybean proteins within the SA network. Not surprisingly, the soybean proteins were from families identified in the literature as being significant contributors to allergens including those identified as belonging to the cupin family: Glyma03g32030.1, Glyma10g39170.1, Glyma03g32020.2, Glyma06g43640.1, Glyma20g28650 (.1., 2), Glyma13g18450.1, Glyma20g28660.1, and Glyma10g39150.1, and to the pathogenesis-related protein Bet v I family: Glyma17g03350 (.1., 2., 3., 4) and Glyma17g03360.1.

The two soybean proteins identified to have high betweenness centralities within the HAR network were also identified as sharing interactions with several curated human proteins within the HAR network. Of these, five were identified among the top 30 by betweenness centrality from the HAR network (Table 7), including P40763 (signal transducer and activator of transcription 3), Q9Y6K9 (NF- κ B essential modulator), P15918 (recombination-activating protein 1), P04179 (superoxide dismutase), and P26038 (moesin).

4.2.1. Limitations of Centrality. While useful in the quantification of proteins within a predicted interactome, centrality measures are not without their limitations. Foremost, no single measure of centrality can be deemed superior to the others. Instead, based on their definitions, each may be favored for a given context over the others. Second, it is a commonly held fallacy that the centrality indicates the relative “importance” of nodes (e.g., $c_b(p_1) = 2 \cdot c_b(p_2)$ implies p_1 is twice as important as p_2) when in fact centrality measures are

explicitly designed to produce a ranking. From a sorted set of centralities, we obtain some indication of the most important nodes. In the context of this work, we sorted each protein by its betweenness centrality to obtain a top- k ranking; however, do not consider the magnitude of the differences between sorted centrality ranks useful. Finally, each centrality measure assumes that each edge shares equal importance and that the network does not contain redundancies (e.g., elimination of a node cannot be compensated by the remaining network). Clearly, in a PPI network, certain edges are far more critical to certain biological processes than others and different biological mechanisms can be compensated by others. Thus, the interpretation of ranked protein centrality should be considered in the context of this naïveté and considered to be potentially subject to noise. Attempts to resolve these limitations by attributing a weighting factor to a given PPI based on some biological measure of importance warrant future investigations. However, this risks introducing a subjective component or being limited by the numerical definition of “importance”.

4.2.2. Sizable Clusters with Degree = 1. The outstanding feature of the predicted networks of Figure 3 is the observation that very large clusters of soybean proteins are predicted to have only a single interaction with one of the HAR proteins of interest. It is imperative to realize that within this network, only interactions involving one of the HAR proteins are depicted; this graph is effectively a *cut* of the completely predicted interactome, which involve predicted interaction across each organism’s complete proteome.

From graph theory, a *cut* is a partition of the vertices of a graph into two disjoint subsets where the edges that have one end point in each subset of the partition comprise the *cut-set* (the edges are said to “cross the cut”). The cut-set in this case comprises all positively predicted PPIs involving the soybean proteins predicted to interact with at least one HAR protein and every other protein in the human proteome. As such, these networks must be interpreted within this isolating context. Interestingly, clusters of various sizes are formed with respect to a number of the HAR proteins, suggestive of the fact that proteins within these clusters might be associated with similar biological functions and/or molecular pathways.

A closer look at the organization of these clusters in the HAR network draws a number of qualitative observations. Discernible single clusters of various sizes of soybean proteins are predicted to interact with each of A0A0A0MT76 (immunoglobulin lambda joining 1), P02768 (serum albumin), O00626 (C–C motif chemokine 22), and, more modestly, Q99650 (oncostatin-M-specific receptor subunit β). Additionally, a number of soybean proteins appear to share interactions across multiple clusters, namely, Q04206 (transcription factor p65), P40763, (signal transducer and activator of transcription 3), P06733 (α -enolase), Glyma01g39580.1 (SUMO conjugation enzyme 1), Glyma02g15310.1 (uncharacterized), Glyma05g17900.1 (SUMO conjugation enzyme 1), Glyma11g05670 (.1., 2., 3., 4) (SUMO conjugation enzyme 1), Glyma16g17410.1 (uncharacterized), Glyma17g18570.1 (SUMO conjugation enzyme 1), and Glyma18g48660.1 (uncharacterized). As hypothesized, proteins present within the same clusters may also be associated with similar biological functions and/or molecular pathways. This is supported as all six soybean proteins have been identified among the top-10 soybean proteins predicted to interact with the HAR by betweenness centrality (Table 9), while Glyma11g05670

(SUMO conjugation enzyme 1) and Glyma18g48660.1 (uncharacterized) have also been identified as the only two proteins from the *G. max* proteome to have large betweenness centralities ranked among the top 30 in the predicted HAR network (Table 7). Among the soy proteins identified as sharing an interaction with all three of the previously mentioned HAR proteins, Glyma01G39580.1 (SUMO conjugation enzyme 1), Glyma05G17900.1 (SUMO conjugation enzyme 1), Glyma11G05670 (.1., 2., 3., 4) (SUMO conjugation enzyme 1), and Glyma17G18570.1 (SUMO conjugation enzyme 1) were all identified as SUMOylating proteins within the soybean proteome, while identified as sharing the same ortholog in *Arabidopsis*, AT3G57870: a SUMO conjugation enzyme 1. With the literature correlating SUMOylation of many proteins, among which are implicated in neurological illnesses including HD, AD, and PD,⁵³ these proteins are of interest for further study. While the family/function of the three remaining proteins, Glyma02G15310.1 (uncharacterized), Glyma16G17410.1 (uncharacterized), and Glyma18G48660.1 (uncharacterized), have yet to be identified in soybean and in *Arabidopsis*, further studies should be conducted.

4.3. Gene Ontology Term Analyses Reveal Pathways Adversely Consequential to Human Health

Investigation of the GO terms of the predicted interactions of the HAR and SA networks revealed a number of biological mechanisms adversely consequential to human health. While the donut charts in Figure 5 only depict the highest level of the GO term hierarchy, the complete set of all ontologies is listed in ST1–ST6.

Most notably, the biological processes related to the immune response were found in both networks. Furthermore, examination of the pathways reveals a large representation of proteins associated with the immune response including the integrin signaling pathway, the TGF- β pathway, and the inflammation mediated by chemokine and cytokine signaling pathways. The TGF- β superfamily of growth factors is composed of several subfamilies of highly conserved cytokines, including TGF- β s.⁵¹ Upon TGF- β signaling, inactive SMAD2 and SMAD3 become phosphorylated via the canonical pathway, together forming a heterodimer.⁵¹ The SMAD2/3 heterodimer then binds to SMAD4 cofactor forming a heterotrimer, facilitating entry into the nucleus and transcriptional activation of target genes.

Interestingly, several pathways related to neurological diseases are also represented including the PD-, HD-, and AD-presenilin pathways. Impairments within the neuronal TGF- β signaling pathway are proposed to play a role in AD pathogenesis including microglia activation, amyloid- β (A β) accumulation, and neurodegeneration.^{51,61–63} In patients with AD and PD, microglia activation resulting in inflammation-related neuronal injuries constitutes a key factor in disease progression and the severity of symptoms.^{64,65} While the precise mechanisms remain unknown, A β accumulation induces synapse loss, apoptosis, and inevitably neurodegeneration along with the associated disease symptoms.⁵¹ Studies conducted in both humans and mouse report that exogenous TGF- β 1 induces microglia activation and removal of A β , suggesting conflicting roles in both apoptosis and survival.^{51,66–69}

Loss of TGF- β signaling regulation mechanisms is apparent in HD pathogenesis.^{51,70–74} Transcriptome comparison

analyses revealed that TGF- β signaling pathway activity increases in instances of HD. This is experimentally shown in induced pluripotent stem cells (iPSCs) and neuronal stem cells from HD patients, as well as *HTT* mutant striatal cell lines and neural progenitor cells derived from iPSCs.^{72,74} In particular, circulating TGF- β 1 exists in increased amounts in symptomatic HD patients and mouse models compared to their asymptomatic counterparts.^{70,71} SMAD7, a negative regulator of TGF- β signaling, was present in significantly lower amounts in *HTT* mutant cell lines, perpetuating the increase of TGF- β signaling in HD neuronal cells.⁷⁴ It remains unclear whether the increase in TGF- β signaling is responsible for neurodegeneration in HD patients or the reciprocal.^{73,74}

In the HAR network, we predict interactions involving four of the isoforms of the 14-3-3 protein (tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein), which have been found to be highly expressed in the brain and have been associated with PD.⁷⁵ Glyma20g28650.2 (cupin family protein) was predicted to interact with four of these isoforms, 14-3-3 η (Q04917), 14-3-3 ζ/δ (P63104), 14-3-3 γ (P61981), and 14-3-3 θ (P27348), in addition to 13 other human proteins. The 14-3-3 proteins are a family of homologous proteins that consist of seven isoforms in mammals.

The four predicted variants are considered to be adapter proteins implicated in the regulation of a large spectrum of both general and specialized signaling pathways. The latter two, γ and θ , have been identified in the literature as being associated with neurological diseases. The 14-3-3 γ protein contributed to early infantile epileptic encephalopathy (EIEE56).⁷⁶ The 14-3-3 θ can be found to be abundantly expressed in the brain and was found to be upregulated in the lumbar spinal cord of patients suffering from sporadic amyotrophic lateral sclerosis (ALS) compared with controls; the highest levels of expression were found in patients with predominant lower motor neuron involvement.⁷⁷ Interestingly, these findings corroborate the findings from the top-10 predictions by betweenness centrality: the mothers against decapentaplegic homologues 2 and 3 have previously been reported to positively regulate 3-phosphoinositide-dependent protein kinase 1 (PDPK1) by stimulating its dissociation from the 14-3-3 θ protein, which functions as a negative regulation factor. PDPK1 has additionally been reported to negatively regulate TGF- β -induced signaling through a number of SMAD proteins.⁷⁸ Both the phosphorylation of the SMAD2/3 heterodimer and nuclear SMAD2, 3, and 4 are decreased in patients with AD, indicating impairment in the TGF- β signaling pathway.^{51,79,80} SMAD7, a known antagonist of TGF- β signaling, has been measured in higher amounts in model mice with familial AD and likely plays a role in apoptosis.^{51,81}

SA interactions with HAR involved with these immune signaling pathways suggest putative targets for the treatment of soybean allergies. PANTHER analysis indicates that a number of SMAD proteins are liable for these terms, including P84022 and Q15796 (SMAD3 and SMAD2, respectively) (ST1–6). P84022 and Q15796 have particularly high betweenness centrality with predicted SA based on the data presented here (SF2). SMAD heteroproteins seek conserved binding sequences called CAGA boxes within promoter regions of target genes.⁸² SMAD3 has been reported to activate the transcription of the *HTT* gene by binding within the promoter region.⁷⁴ Disruptions to the SMAD heteroprotein complex

formation or binding to conserved CAGA sequences serve as targets for potential soybean allergy therapeutics. Continued research in this area should be explored.

4.4. Insights into Proteins Sharing Similar Interaction Profiles

The rank ordering of the similarity profiles produces a number of interesting distributions and properties. Traditional approaches might consider taking the top- k of proteins based on their sorted Sørensen–Dice coefficient or threshold the results by some globally applied value, τ . This approach assumes that each protein has a similar distribution of shared highly similar interaction profiles. In this work, we leverage the one-to-all score curve inspired from the reciprocal perspective framework^{42,43} to visualize the rank-order distribution of similarity coefficients for each protein to better understand the relative similarity between any protein relative to all others in the proteome. A protein sharing a large degree of similarity with a very large proportion of proteins in the proteome might be considered less interesting than a protein that shares a large degree of similarity with only a few rare proteins.

From a small, curated subset of nine proteins, using both one-to-all score curves and violin plots to highlight the variation in similarity coefficient densities, we immediately notice that the soybean proteins have a much larger number of similarly interacting proteins: the majority of the comprehensive interactome displaying the near-linear decreasing or the stepwise decreasing behavior while the human proteins generally displaying far fewer (Figure 6). This is not surprising given that soybean has undergone two whole-genome duplications. Furthermore, the stepwise distribution of similarity coefficient is indicative of a *small interaction set effect* where the progressive elimination of single shared interactions results in a large reduction in the similarity coefficient. Specifically, numerous proteins with the same sized set of interactions and sharing this same small network result in a plateaued distribution, and as shared nodes are progressively dropped, we step down to the next plateau of similarity. In theory, the number of these distinct steps can depict the size of the comparative network. As a given protein's predicted network size increases, the granularity of these steps becomes progressively finer; proteins with a modest similarity coefficient might fully share another protein's set of interactors; however, they may also have a large number of other interactors that progressively decreases the similarity coefficient.

At its most extreme, the near-linearly decreasing behavior seen in a subset of the soybean proteins suggests that within the comprehensive interactome there exists a large central interconnected network, which the majority of the soybean proteome is predicted to interact with. This is not necessarily surprising since these biological interactomes are often modeled as *small-world* networks having a high clustering coefficient. Contrastingly, a minimal number of human proteins appear to display the near-linearly decreasing behavior; rather, the L-shaped distribution is found to be represented by the majority of human proteins. This suggests that within the comprehensive interactome there is a rather small, interconnected network to which the different subsets within the human proteome are predicted to interact with.

By leveraging the hypothesis originated from the TCP, putative allergens may be identified based on the similarity of function and/or pathway associated with the HAR and SA networks. With a small subset shown in Figure 6, it was found

that only a small number of proteins are found to share a high degree of similarity with the comprehensive interactome, relative to all other proteins in the proteome. The most extreme representation of this behavior is described as the L-shaped distribution, with only a few rare proteins found to share any above-zero similarity, while the remainder of the proteome shares none. While qualitatively determined, this interaction profile was found to only be visible in a small subset of the comprehensive interactome that includes two profilin-1 proteins Glyma08G03660 (profilin 5) and Glyma09G07800 (profilin) and two uncharacterized proteins (Glyma10G09890 and Glyma18G32740 (uncharacterized)) suggestive of further investigation of potential putative proteins attributed to human health among the respective interactomes. Rather, only one protein, Glyma08G03660 (profilin 5), was identified as having a significant interaction profile with a high Sørensen coefficient with a protein within the proteome that has not already been identified as being within the comprehensive interactome Glyma15G02380.1 (profilin 4).

In contrast to what was observed among the soybean comprehensive interactome, the human proteins that were predicted within the SA network were found to not exist within a large, interconnected network, in part, due to the absence of proteins identified as having a linear decreasing interaction profile and minimal proteins exemplifying the stepwise interaction profile. Rather, the majority of proteins within this comprehensive interactome were identified as having a small, interconnected network, with most identified as sharing the L-shaped interaction profile. Among the proteins previously mentioned to have the extreme form of the L-shaped distribution, each with a small, nonzero similarity shared across thousands of proteins in the proteome, were P01308 (insulin), P56537 (eukaryotic translation initiation factor 6), and P51671 (eotaxin). Q99616 (C–C chemokine 13) was identified as the only protein with a significant similarity coefficient to a known allergen Q619T4 (C–C motif chemokine 11). Finally, members of the chemokine family, P13500 (C–C chemokine 2) and P80098 (C–C chemokine 7), have been implicated in allergic lung inflammation.

5. CONCLUSIONS

In this work, by investigating the interacting partners of curated proteins associated with the human allergy response (HAR) and soybean allergen (SA), we were able to identify *de novo* PPIs consequential to human health. A total of eight proteins with significant protein connectivity in contrasting networks were identified, composed of six from the human proteome and two from the soybean proteome, supported by the literature to play a role in immunological and neurological pathways. Our approach involved the use of the protein–protein interaction prediction engine (PIPE4) algorithm, which was specifically developed for complex inter- and cross-species prediction schemas to generate the comprehensive interactome between *H. sapiens* and *G. max*. A literature-curated list of human proteins known to be associated with HAR and a second literature-curated list of soybean proteins known to be associated with SA were used to identify candidate proteins whose interactions may be consequential to human health. This study, beyond generating the most comprehensive human–soybean interactome to date, elucidated a soybean seed interactome and identified several seed proteins significant to human health.

■ ASSOCIATED CONTENT

SI Supporting Information

The Supporting Information is available free of charge at <https://pubs.acs.org/doi/10.1021/acs.jproteome.1c00138>.

Predicted *G. max* biological process over-representation test (Table S1); predicted *G. max* molecular function over-representation test (Table S2); predicted *G. max* cellular component over-representation test (Table S3); predicted *H. sapiens* biological process over-representation test (Table S4); predicted *H. sapiens* molecular function over-representation test (Table S5); predicted *H. sapiens* cellular component over-representation test (Table S6); soy-seed intersected interactome (Table S7); human-seed intersected interactome (Table S8); literature findings of immunological approaches that validate the key protein involvement in human health (Table S9); known allergens from other species related to soybean (Table S10); all-to-all predictions between legume proteins and human targets (Table S11); the predicted interactome for proteins known to be involved in the human allergy response (Figure S1); the predicted interactome for known soybean allergens (Figure S2); and the predicted interactome for proteins believed to be among the soybean seed proteome (Figure S3) (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Bahram Samanfar – Agriculture and Agri-Food Canada, Ottawa Research and Development Centre, Ottawa, Ontario, Canada K1A 0C6; Department of Biology and Institute of Biochemistry, and Ottawa Institute of Systems Biology, Carleton University, Ottawa, Ontario, Canada K1S 5B6; orcid.org/0000-0001-6625-2153; Email: bahram.samanfar@agr.gc.ca

Authors

Kevin Dick – Department of Systems and Computer Engineering, Carleton University, Ottawa, Ontario, Canada K1S 5B6

Arezo Pattang – Agriculture and Agri-Food Canada, Ottawa Research and Development Centre, Ottawa, Ontario, Canada K1A 0C6; Department of Biology and Institute of Biochemistry, and Ottawa Institute of Systems Biology, Carleton University, Ottawa, Ontario, Canada K1S 5B6

Julia Hooker – Agriculture and Agri-Food Canada, Ottawa Research and Development Centre, Ottawa, Ontario, Canada K1A 0C6; Department of Biology and Institute of Biochemistry, and Ottawa Institute of Systems Biology, Carleton University, Ottawa, Ontario, Canada K1S 5B6

Nour Nissan – Agriculture and Agri-Food Canada, Ottawa Research and Development Centre, Ottawa, Ontario, Canada K1A 0C6; Department of Biology and Institute of Biochemistry, and Ottawa Institute of Systems Biology, Carleton University, Ottawa, Ontario, Canada K1S 5B6

Michael Sadowski – Agriculture and Agri-Food Canada, Ottawa Research and Development Centre, Ottawa, Ontario, Canada K1A 0C6; Department of Biology and Institute of Biochemistry, and Ottawa Institute of Systems Biology, Carleton University, Ottawa, Ontario, Canada K1S 5B6

Bradley Barnes – Department of Systems and Computer Engineering, Carleton University, Ottawa, Ontario, Canada K1S 5B6

Le Hoa Tan – Agriculture and Agri-Food Canada, Ottawa Research and Development Centre, Ottawa, Ontario, Canada K1A 0C6; Department of Biology and Institute of Biochemistry, and Ottawa Institute of Systems Biology, Carleton University, Ottawa, Ontario, Canada K1S 5B6

Daniel Burnside – Department of Biology and Institute of Biochemistry, and Ottawa Institute of Systems Biology, Carleton University, Ottawa, Ontario, Canada K1S 5B6

Sadhna Phanse – Department of Biochemistry, University of Regina, Regina, Saskatchewan, Canada S4S 0A2

Hiroyuki Aoki – Department of Biochemistry, University of Regina, Regina, Saskatchewan, Canada S4S 0A2

Mohan Babu – Department of Biochemistry, University of Regina, Regina, Saskatchewan, Canada S4S 0A2

Frank Dehne – School of Computer Science, Carleton University, Ottawa, Ontario, Canada K1S 5B6

Ashkan Golshani – Department of Biology and Institute of Biochemistry, and Ottawa Institute of Systems Biology, Carleton University, Ottawa, Ontario, Canada K1S 5B6

Elroy R. Cober – Agriculture and Agri-Food Canada, Ottawa Research and Development Centre, Ottawa, Ontario, Canada K1A 0C6

James R. Green – Department of Systems and Computer Engineering, Carleton University, Ottawa, Ontario, Canada K1S 5B6

Complete contact information is available at:

<https://pubs.acs.org/doi/10.1021/acs.jproteome.1c00138>

Notes

The authors declare no competing financial interest.

#J.H., N.N., and A.P. contributed equally to this work.

■ ACKNOWLEDGMENTS

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE^{83,84} partner repository with the data set identifier PXD023875. This work has been dedicated to the loving memory of Fareed Arasteh, one of the passengers of the Ukrainian airlines flight 752.

■ REFERENCES

- (1) Karuga, J. 10 Countries With Largest Soybean Production.
- (2) Cordle, C. T. Soy Protein Allergy: Incidence and Relative Severity. *J. Nutr.* **2004**, *134*, 1213S–1219S.
- (3) World Health Organization and International Union of Immunological Societies Allergen Nomenclature Sub-Committee. WHO/IUIS Allergen Nomenclature Sub-Committee.
- (4) Awazuhara, H.; Kawai, H.; Maruchi, N. Major Allergens in Soybean and Clinical Significance of IgG4 Antibodies Investigated by IgE- and IgG4-Immunoblotting with Sera from Soybean-Sensitive Patients. *Clin. Exp. Allergy* **1997**, *27*, 325–332.
- (5) Batista, R.; Martins, I.; Jenó, P.; Ricardo, C. P.; Oliveira, M. M. A Proteomic Study to Identify Soya Allergens - The Human Response to Transgenic versus Non-Transgenic Soya Samples. *Int. Arch. Allergy Immunol.* **2007**, *144*, 29–38.
- (6) Gómez-Ollés, S.; Cruz, M. J.; Bogdanovic, J.; Wouters, I. M.; Doekes, G.; Sander, I.; Morell, F.; Rodrigo, M. J. Assessment of Soy Aeroallergen Levels in Different Work Environments. *Clin. Exp. Allergy* **2007**, *37*, 1863–1872.
- (7) Fukushima, D. Soy Proteins. In *Handbook of Food Proteins*; Woodhead Pub: Cambridge, 2011; pp 210–232.

- (8) Herian, A. M.; Taylor, S. L.; Bush, R. K. Allergenic Reactivity of Various Soybean Products as Determined by RAST Inhibition. *J. Food Sci.* **1993**, *58*, 385–388.
- (9) Nielsen, N. C.; Dickinson, C. D.; Cho, T. J.; Thanh, V. H.; Scallan, B. J.; Fischer, R. L.; Sims, T. L.; Drews, G. N.; Goldberg, R. B. Characterization of the Glycinin Gene Family in Soybean. *Plant Cell* **1989**, *1*, 313–328.
- (10) Ogawa, T.; Bando, N.; Tuji, H.; Nishikawa, K.; Kitamura, K. α -Subunit of β -Conglycinin, an Allergenic Protein Recognized by IgE Antibodies of Soybean-Sensitive Patients with Atopic Dermatitis. *Biosci., Biotechnol., Biochem.* **1995**, *59*, 831–833.
- (11) Ogawa, T.; Samoto, M.; Takahashi, K. Soybean Allergens and Hypoallergenic Soybean Products. *J. Nutr. Sci. Vitaminol.* **2000**, *46*, 271–279.
- (12) Herman, E. M. Genetically Modified Soybeans and Food Allergies. *J. Exp. Bot.* **2003**, *54*, 1317–1319.
- (13) Mills, E. N. C.; Jenkins, J.; Marigheto, N.; Belton, P. S.; Gunning, A. P.; Morris, V. J. Allergens of the Cupin Superfamily. In *Biochemical Society Transactions*; Portland Press, 2002; Vol. 30, pp 925–929.
- (14) Krishnan, H. B. Biochemistry and Molecular Biology of Soybean Seed Storage Proteins. *J. New Seeds* **2000**, *2*, 1–25.
- (15) Herman, E. M. Soybean Seed Proteome Rebalancing. *Front. Plant Sci.* **2014**, *5*, 437.
- (16) Krishnan, H. B.; Kim, W. S.; Jang, S.; Kerley, M. S. All Three Subunits of Soybean β -Conglycinin Are Potential Food Allergens. *J. Agric. Food Chem.* **2009**, *57*, 938–943.
- (17) Breiteneder, H.; Radauer, C. A Classification of Plant Food Allergens. *J. Allergy Clin. Immunol.* **2004**, *113*, 821–830.
- (18) Xu, J.; Li, Y. Discovering Disease-Genes by Topological Features in Human Protein–Protein Interaction Network. *Bioinformatics* **2006**, *22*, 2800–2805.
- (19) Zhao, B.; Wang, J.; Li, M.; Li, X.; Li, Y.; Wu, F.-X.; Pan, Y. A New Method for Predicting Protein Functions From Dynamic Weighted Interactome Networks. *IEEE Trans. NanoBiosci.* **2016**, *15*, 131–139.
- (20) Gligorijević, V.; Barot, M.; Bonneau, R.; Wren, J. DeepNF: Deep Network Fusion for Protein Function Prediction. *Bioinformatics* **2018**, *34*, 3873–3881.
- (21) Yildirim, M. A.; Goh, K.-I.; Cusick, M. E.; Barabási, A.-L.; Vidal, M. Drug–Target Network. *Nat. Biotechnol.* **2007**, *25*, 1119–1126.
- (22) Schoenrock, A.; Burnside, D.; Moteshareie, H.; Wong, A.; Golshani, A.; Dehne, F.; Green, J. R. Engineering Inhibitory Proteins with InSiPS: The in-Silico Protein Synthesizer, In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis on - SC '15*, ACM Press: New York, New York, USA, 2015; pp 1–11.
- (23) Georg, S. *Soziologie: Untersuchungen Über Die Formen Der Vergesellschaftung*; Duncker & Humboldt: Leipzig, 1908.
- (24) Granovetter, M. S. The Strength of Weak Ties. *Am. J. Sociol.* **1973**, *78*, 1360–1380.
- (25) Cook, H.; Doncheva, N.; Szklarczyk, D.; von Mering, C.; Jensen, L. Viruses.STRING: A Virus-Host Protein-Protein Interaction Database. *Viruses* **2018**, *10*, 519.
- (26) Zhang, Q. C.; Petrey, D.; Garzon, J. I.; Deng, L.; Honig, B. PrePPI: A Structure-Informed Database of Protein-Protein Interactions. *Nucleic Acids Res.* **2012**, *41*, D828–D833.
- (27) Schoenrock, A.; Dehne, F.; Green, J. R.; Golshani, A.; Pitre, S. MP-PIPE: A Massively Parallel Protein-Protein Interaction Prediction Engine. In *Proceedings of the international conference on Supercomputing - ICS '11*; ACM Press: Tucson, Arizona, USA, 2011; p 327.
- (28) Li, Y.; Ilie, L. SPRINT: Ultrafast Protein-Protein Interaction Prediction of the Entire Human Interactome. *BMC Bioinf.* **2017**, *18*, No. 485.
- (29) Dick, K.; Dehne, F.; Golshani, A.; Green, J. R. Positome: A Method for Improving Protein-Protein Interaction Quality and Prediction Accuracy. In *2017 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*; IEEE, 2017; pp 1–8.
- (30) Pitre, S.; Hooshyar, M.; Schoenrock, A.; Samanfar, B.; Jessulat, M.; Green, J. R.; Dehne, F.; Golshani, A. Short Co-Occurring Polypeptide Regions Can Predict Global Protein Interaction Maps. *Sci. Rep.* **2012**, *2*, No. 239.
- (31) Dick, K.; Samanfar, B.; Barnes, B.; Cober, E. R.; Mimeo, B.; Tan, L. H.; Molnar, S. J.; Biggar, K. K.; Golshani, A.; Dehne, F.; Green, J. R. PIPE4: Fast PPI Predictor for Comprehensive Inter- and Cross-Species Interactomes. *Sci. Rep.* **2020**, *10*, No. 1390.
- (32) Girvan, M.; Newman, M. E. J. Community Structure in Social and Biological Networks. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 7821–7826.
- (33) Pitre, S.; Dehne, F.; Chan, A.; Cheetham, J.; Duong, A.; Emili, A.; Gebbia, M.; Greenblatt, J.; Jessulat, M.; Krogan, N.; Luo, X.; Golshani, A. PIPE: A Protein-Protein Interaction Prediction Engine Based on the Re-Occurring Short Polypeptide Sequences between Known Interacting Protein Pairs. *BMC Bioinf.* **2006**, *7*, No. 365.
- (34) Pitre, S.; North, C.; Alamgir, M.; Jessulat, M.; Chan, A.; Luo, X.; Green, J. R.; Dumontier, M.; Dehne, F.; Golshani, A. Global Investigation of Protein–Protein Interactions in Yeast *Saccharomyces Cerevisiae* Using Re-Occurring Short Polypeptide Sequences. *Nucleic Acids Res.* **2008**, *36*, 4286–4294.
- (35) Schoenrock, A.; Samanfar, B.; Pitre, S.; Hooshyar, M.; Jin, K.; Phillips, C. A.; Wang, H.; Phanse, S.; Omidi, K.; Gui, Y.; Alamgir, M.; Wong, A.; Barrenäs, F.; Babu, M.; Benson, M.; Langston, M. A.; Green, J. R.; Dehne, F.; Golshani, A. Efficient Prediction of Human Protein-Protein Interactions at a Global Scale. *BMC Bioinf.* **2014**, *15*, No. 383.
- (36) Grant, D.; Nelson, R. T.; Cannon, S. B.; Shoemaker, R. C. SoyBase, the USDA-ARS Soybean Genetics and Genomics Database. *Nucleic Acids Res.* **2010**, *38*, D843–D846.
- (37) Yates, J. R., III; Eng, J. K.; Clauser, K. R.; Burlingame, A. L. Search of Sequence Databases with Uninterpreted High-Energy Collision-Induced Dissociation Spectra of Peptides. *J. Am. Soc. Mass Spectrom.* **1996**, *7*, 1089–1098.
- (38) Kislinger, T.; Rahman, K.; Radulovic, D.; Cox, B.; Rossant, J.; Emili, A. PRISM, a Generic Large Scale Proteomic Investigation Strategy for Mammals. *Mol. Cell. Proteomics* **2003**, *2*, 96–106.
- (39) Sabidussi, G. The Centrality Index of a Graph. *Psychometrika* **1966**, *31*, 581–603.
- (40) Mi, H.; Huang, X.; Muruganujan, A.; Tang, H.; Mills, C.; Kang, D.; Thomas, P. D. PANTHER Version 11: Expanded Annotation Data from Gene Ontology and Reactome Pathways, and Data Analysis Tool Enhancements. *Nucleic Acids Res.* **2017**, *45*, D183–D189.
- (41) Kovács, I. A.; Luck, K.; Spirohn, K.; Wang, Y.; Pollis, C.; Schlabach, S.; Bian, W.; Kim, D.-K.; Kishore, N.; Hao, T.; Calderwood, M. A.; Vidal, M.; Barabási, A.-L. Network-Based Prediction of Protein Interactions. *bioRxiv* **2018**, No. 275529.
- (42) Dick, K.; Green, J. R. Reciprocal Perspective for Improved Protein-Protein Interaction Prediction. *Sci. Rep.* **2018**, *8*, No. 11694.
- (43) Kyrollos, D. G.; Reid, B.; Dick, K.; Green, J. R. RpmirDIP: Reciprocal Perspective Improves MiRNA Targeting Prediction. *Sci. Rep.* **2020**, *10*, No. 11770.
- (44) Taylor, S. L.; Hefle, S. L. Food Allergies and Other Food Sensitivities. *Food Technol.* **2001**, *55*, 68.
- (45) Ogawa, T.; Bando, N.; Tsuji, H.; Okajima, H.; Nishikawa, K.; Sasaoka, K. Investigation of the IgE-Binding Proteins in Soybeans by Immunoblotting with the Sera of the Soybean-Sensitive Patients with Atopic Dermatitis. *J. Nutr. Sci. Vitaminol.* **1991**, *37*, 555–565.
- (46) L'Hocine, L.; Boye, J. I. Allergenicity of Soybean: New Developments in Identification of Allergenic Proteins, Cross-Reactivities and Hypoallergenization Technologies. *Crit. Rev. Food Sci. Nutr.* **2007**, *47*, 127–143.
- (47) Joy, M. P.; Brock, A.; Ingber, D. E.; Huang, S. High-Betweenness Proteins in the Yeast Protein Interaction Network. *J. Biomed. Biotechnol.* **2005**, *2005*, 96–103.

- (48) Shamsi, T. N.; Parveen, R.; Fatima, S. Characterization, Biomedical and Agricultural Applications of Protease Inhibitors: A Review. *Int. J. Biol. Macromol.* **2016**, *91*, 1120–1133.
- (49) Mullan, M.; Crawford, F.; Axelman, K.; Houlden, H.; Lilius, L.; Winblad, B.; Lannfelt, L. A Pathogenic Mutation for Probable Alzheimer's Disease in the APP Gene at the N-Terminus of β -Amyloid. *Nat. Genet.* **1992**, *1*, 345–347.
- (50) Murrell, J. R.; Hake, A. M.; Quaid, K. A.; Farlow, M. R.; Ghetti, B. Early-Onset Alzheimer Disease Caused by a New Mutation (V717L) in the Amyloid Precursor Protein Gene. *Arch. Neurol.* **2000**, *57*, 885.
- (51) Kashima, R.; Hata, A. The Role of TGF- β Superfamily Signaling in Neurological Disorders. *Acta Biochim. Biophys. Sin.* **2018**, *50*, 106–120.
- (52) Lee, L.; Dale, E.; Staniszewski, A.; Zhang, H.; Saeed, F.; Sakurai, M.; Fa', M.; Orozco, I.; Michelassi, F.; Akpan, N.; Lehrer, H.; Arancio, O. Regulation of Synaptic Plasticity and Cognition by SUMO in Normal Physiology and Alzheimer's Disease. *Sci. Rep.* **2014**, *4*, No. 7190.
- (53) Sarge, K. D.; Park-Sarge, O.-K. Sumoylation and Human Disease Pathogenesis. *Trends Biochem. Sci.* **2009**, *34*, 200–205.
- (54) Bergink, S.; Jentsch, S. Principles of Ubiquitin and SUMO Modifications in DNA Repair. *Nature* **2009**, *458*, 461–467.
- (55) Mukhopadhyay, D.; Dasso, M. Modification in Reverse: The SUMO Proteases. *Trends Biochem. Sci.* **2007**, *32*, 286–295.
- (56) Anderson, D. B.; Wilkinson, K. A.; Henley, J. M. Protein Sumoylation In Neuropathological Conditions. *Drug News Perspect.* **2009**, *22*, 255–265.
- (57) Yang, W.; Sheng, H.; Wang, H. Targeting the SUMO Pathway for Neuroprotection in Brain Ischaemia. *Stroke Vasc. Neurol.* **2016**, *1*, 101–107.
- (58) Marlier, Q.; Jibassia, F.; Verteneuil, S.; Linden, J.; Kaldis, P.; Meijer, L.; Nguyen, L.; Vandenbosch, R.; Malgrange, B. Genetic and Pharmacological Inhibition of Cdk1 Provides Neuroprotection towards Ischemic Neuronal Death. *Cell Death Discovery* **2018**, *4*, 1–2.
- (59) Park, D. S.; Levine, B.; Ferrari, G.; Greene, L. A. Cyclin Dependent Kinase Inhibitors and Dominant Negative Cyclin Dependent Kinase 4 and 6 Promote Survival of NGF-Deprived Sympathetic Neurons. *J. Neurosci.* **1997**, *17*, 8975–8983.
- (60) Mergenthaler, P.; Muselmann, C.; Sünwoldt, J.; Isaev, N. K.; Wieloch, T.; Dirnagl, U.; Meisel, A.; Ruscher, K. A Functional Role of the Cyclin-Dependent Kinase Inhibitor 1 (P21 WAF1/CIP1) for Neuronal Preconditioning. *J. Cereb. Blood Flow Metab.* **2013**, *33*, 351–355.
- (61) Lauzon, M. A.; Daviau, A.; Marcos, B.; Faucheux, N. Nanoparticle-Mediated Growth Factor Delivery Systems: A New Way to Treat Alzheimer's Disease. *J. Controlled Release* **2015**, *206*, 187–205.
- (62) Tesseur, I.; Zou, K.; Esposito, L.; Bard, F.; Berber, E.; Van Can, J.; Lin, A. H.; Crews, L.; Tremblay, P.; Mathews, P.; Mucke, L.; Masliah, E.; Wyss-Coray, T. Deficiency in Neuronal TGF- β Signaling Promotes Neurodegeneration and Alzheimer's Pathology. *J. Clin. Invest.* **2006**, *116*, 3060–3069.
- (63) Tichauer, J. E.; von Bernhardi, R. Transforming Growth Factor- β Stimulates β Amyloid Uptake by Microglia through Smad3-Dependent Mechanisms. *J. Neurosci. Res.* **2012**, *90*, 1970–1980.
- (64) Thameem Dheen, S.; Kaur, C.; Ling, E.-A. Microglial Activation and Its Implications in the Brain Diseases. *Curr. Med. Chem.* **2007**, *14*, 1189–1197.
- (65) Bachiller, S.; Jiménez-Ferrer, I.; Paulus, A.; Yang, Y.; Swanberg, M.; Deierborg, T.; Boza-Serrano, A. Microglia in Neurological Diseases: A Road Map to Brain-Disease Dependent-Inflammatory Response. *Front. Cell. Neurosci.* **2018**, *12*, 488.
- (66) Wyss-Coray, T.; Lin, C.; Yan, F.; Yu, G. Q.; Rohde, M.; McConlogue, L.; Masliah, E.; Mucke, L. TGF- β 1 Promotes Microglial Amyloid-Beta Clearance and Reduces Plaque Burden in Transgenic Mice. *Nat. Med.* **2001**, *7*, 612–618.
- (67) Caraci, F.; Battaglia, G.; Bruno, V.; Bosco, P.; Carbonaro, V.; Giuffrida, M. L.; Drago, F.; Sortino, M. A.; Nicoletti, F.; Copani, A. TGF- β 1 Pathway as a New Target for Neuroprotection in Alzheimer's Disease. *CNS Neurosci. Ther.* **2011**, *17*, 237–249.
- (68) Fisichella, V.; Giurdanella, G.; Platania, C. B. M.; Romano, G. L.; Leggio, G. M.; Salomone, S.; Drago, F.; Caraci, F.; Bucolo, C. TGF- β 1 Prevents Rat Retinal Insult Induced by Amyloid- β (1–42) Oligomers. *Eur. J. Pharmacol.* **2016**, *787*, 72–77.
- (69) Diniz, L. P.; Tortelli, V.; Matias, I.; Morgado, J.; Bérnago Araujo, A. P.; Melo, H. M.; Seixas da Silva, G. S.; Alves-Leon, S. V.; de Souza, J. M.; Ferreira, S. T.; De Felice, F. G.; Gomes, F. C. A. Astrocyte Transforming Growth Factor Beta 1 Protects Synapses against A β Oligomers in Alzheimer's Disease Model. *J. Neurosci.* **2017**, *37*, 6797–6809.
- (70) Chang, K.-H.; Wu, Y.-R.; Chen, Y.-C.; Chen, C.-M. Plasma Inflammatory Biomarkers for Huntington's Disease Patients and Mouse Model. *Brain, Behav., Immun.* **2015**, *44*, 121–127.
- (71) Battaglia, G.; Cannella, M.; Rizzo, B.; Orobello, S.; Maat-Schieman, M. L.; Aronica, E.; Busceti, C. L.; Ciarmiello, A.; Alberti, S.; Amico, E.; Sassone, J.; Sipione, S.; Bruno, V.; Frati, L.; Nicoletti, F.; Squitieri, F. Early Defect of Transforming Growth Factor B1 Formation in Huntington's Disease. *J. Cell. Mol. Med.* **2011**, *15*, 555–571.
- (72) Ring, K. L.; An, M. C.; Zhang, N.; O'Brien, R. N.; Ramos, E. M.; Gao, F.; Atwood, R.; Bailus, B. J.; Melov, S.; Mooney, S. D.; Coppola, G.; Ellerby, L. M. Genomic Analysis Reveals Disruption of Striatal Neuronal Development and Therapeutic Targets in Human Huntington's Disease Neural Stem Cells. *Stem Cell Rep.* **2015**, *5*, 1023–1038.
- (73) Kandasamy, M.; Couillard-Despres, S.; Raber, K. A.; Stephan, M.; Lehner, B.; Winner, B.; Kohl, Z.; Rivera, F. J.; Nguyen, H. P.; Riess, O.; Bogdahn, U.; Winkler, J.; von Hörsten, S.; Aigner, L. Stem Cell Quiescence in the Hippocampal Neurogenic Niche Is Associated with Elevated Transforming Growth Factor-Beta Signaling in an Animal Model of Huntington Disease. *J. Neuropathol. Exp. Neurol.* **2010**, *69*, 717–728.
- (74) Bowles, K. R.; Stone, T.; Holmans, P.; Allen, N. D.; Dunnett, S. B.; Jones, L. SMAD Transcription Factors Are Altered in Cell Models of HD and Regulate HTT Expression. *Cell. Signalling* **2017**, *31*, 1–14.
- (75) Foote, M.; Zhou, Y. 14-3-3 Proteins in Neurological Disorders. *Int. J. Biochem. Mol. Biol.* **2012**, *3*, 152–164.
- (76) Guella, I.; McKenzie, M. B.; Evans, D. M.; Buerki, S. E.; Toyota, E. B.; Van Allen, M. I.; Adam, S.; Boelman, C.; Bolbocean, C.; Candido, T.; Eydoux, P.; Horvath, G.; Huh, L.; Nelson, T. N.; Sinclair, G.; van Karnebeek, C.; Vercauteren, S.; Suri, M.; Elmslie, F.; Simon, M. E. H.; van Gassen, K. L. I.; Héron, D.; Keren, B.; Nava, C.; Connolly, M. B.; Demos, M.; Farrer, M. J. De Novo Mutations in YWHAG Cause Early-Onset Epilepsy. *Am. J. Hum. Genet.* **2017**, *101*, 300–310.
- (77) Malaspina, A.; Kaushik, N.; Bellerocche, J. de. A 14-3-3 MRNA Is Up-Regulated in Amyotrophic Lateral Sclerosis Spinal Cord. *J. Neurochem.* **2000**, *75*, 2511–2520.
- (78) Seong, H.-A.; Jung, H.; Kim, K.-T.; Ha, H. 3-Phosphoinositide-Dependent PDK1 Negatively Regulates Transforming Growth Factor-Beta-Induced Signaling in a Kinase-Dependent Manner through Physical Interaction with Smad Proteins. *J. Biol. Chem.* **2007**, *282*, 12272–12289.
- (79) Ueberham, U.; Ueberham, E.; Gruschka, H.; Arendt, T. Altered Subcellular Location of Phosphorylated Smads in Alzheimer's Disease. *Eur. J. Neurosci.* **2006**, *24*, 2327–2334.
- (80) Ueberham, U.; Hilbrich, I.; Ueberham, E.; Rohn, S.; Glöckner, P.; Dietrich, K.; Brückner, M. K.; Arendt, T. Transcriptional Control of Cell Cycle-Dependent Kinase 4 by Smad Proteins—Implications for Alzheimer's Disease. *Neurobiol. Aging* **2012**, *33*, 2827–2840.
- (81) Salins, P.; He, Y.; Olson, K.; Glazner, G.; Kashour, T.; Amara, F. TGF- β 1 Is Increased in a Transgenic Mouse Model of Familial Alzheimer's Disease and Causes Neuronal Apoptosis. *Neurosci. Lett.* **2008**, *430*, 81–86.
- (82) Dennler, S.; Itoh, S.; Vivien, D.; Gauthier, J. Direct Binding of Smad3 and Smad4 to Critical TGF β 1-inducible elements in the

promoter of human plasminogen activator inhibitor-type 1 gene. *EMBO J.* **1998**, *17*, 3091–3100.

(83) Perez-Riverol, Y.; Csordas, A.; Bai, J.; Bernal-Llinares, M.; Hewapathirana, S.; Kundu, D. J.; Inuganti, A.; Griss, J.; Mayer, G.; Eisenacher, M.; Pérez, E.; Uszkoreit, J.; Pfeuffer, J.; Sachsenberg, T.; Yilmaz, S.; Tiwary, S.; Cox, J.; Audain, E.; Walzer, M.; Jarnuczak, A. F.; Ternent, T.; Brazma, A.; Vizcaino, J. A. The PRIDE Database and Related Tools and Resources in 2019: Improving Support for Quantification Data. *Nucleic Acids Res.* **2019**, *47*, D442–D450.

(84) Deutsch, E. W.; Bandeira, N.; Sharma, V.; Perez-Riverol, Y.; Carver, J. J.; Kundu, D. J.; García-Seisdedos, D.; Jarnuczak, A. F.; Hewapathirana, S.; Pullman, B. S.; Wertz, J.; Sun, Z.; Kawano, S.; Okuda, S.; Watanabe, Y.; Hermjakob, H.; Maclean, B.; Maccoss, M. J.; Zhu, Y.; Ishihama, Y.; Vizcaino, J. A. The ProteomeXchange Consortium in 2020: Enabling “big Data” Approaches in Proteomics. *Nucleic Acids Res.* **2020**, *48*, D1145–D1152.

(85) Gonzalez, R.; Polo, F.; Zapatero, L.; Caravaca, F.; Carreira, J. Purification and Characterization of Major Inhalant Allergens from Soybean Hulls. *Clin. Exp. Allergy* **1992**, *22*, 748–755.

(86) Rihs, H. P.; Chen, Z.; Ruëff, F.; Petersen, A.; Rozynek, P.; Heimann, H.; Baur, X. IgE Binding of the Recombinant Allergen Soybean Profilin (RGly m 3) Is Mediated by Conformational Epitopes. *J. Allergy Clin. Immunol.* **1999**, *104*, 1293–1301.

(87) Crowell, D. N.; John, M. E.; Russell, D.; Amasino, R. M. Characterization of a Stress-Induced, Developmentally Regulated Gene Family from Soybean. *Plant Mol. Biol.* **1992**, *18*, 459–466.

(88) Natarajan, S.; Xu, C.; Bae, H.; Bailey, B. A.; Cregan, P.; Caperna, T. J.; Garrett, W. M.; Luthria, D. Proteomic and Genetic Analysis of Glycinin Subunits of Sixteen Soybean Genotypes. *Plant Physiol. Biochem.* **2007**, *45*, 436–444.

(89) Riascos, J. J.; Weissinger, S. M.; Weissinger, A. K.; Kulis, M.; Burks, A. W.; Pons, L. The Seed Biotinylated Protein of Soybean (Glycine Max): A Boiling-Resistant New Allergen (Gly m 7) with the Capacity To Induce IgE-Mediated Allergic Responses. *J. Agric. Food Chem.* **2016**, *64*, 3890–3900.

(90) Ebisawa, M.; Brostedt, P.; Sjölander, S.; Sato, S.; Borres, M. P.; Ito, K. Gly m 2S Albumin Is a Major Allergen with a High Diagnostic Value in Soybean-Allergic Children. *J. Allergy Clin. Immunol.* **2013**, *132*, 976–978.e5.

(91) Fu, T. J.; Abbott, U. R.; Hatzos, C. Digestibility of Food Allergens and Nonallergenic Proteins in Simulated Gastric Fluid and Simulated Intestinal Fluid - A Comparative Study. *J. Agric. Food Chem.* **2002**, *50*, 7154–7160.

(92) Liu, B.; Teng, D.; Wang, X.; Wang, J. Detection of the Soybean Allergenic Protein Gly m Bd 28K by an Indirect Enzyme-Linked Immunosorbent Assay. *J. Agric. Food Chem.* **2013**, *61*, 822–828.

(93) Morita, H.; Kaneko, H.; Ohnishi, H.; Kato, Z.; Kubota, K.; Yamamoto, T.; Matsui, E.; Teramoto, T.; Fukao, T.; Kasahara, K.; Kondo, N. Structural Property of Soybean Protein P34 and Specific IgE Response to Recombinant P34 in Patients with Soybean Allergy. *Int. J. Mol. Med.* **2012**, *29*, 153–158.

(94) Gagnon, C.; Poysa, V.; Cober, E. R.; Gleddie, S. Soybean Allergens Affecting North American Patients Identified by 2D Gels and Mass Spectrometry. *Food Anal. Methods* **2010**, *3*, 363–374.

Supplementary Information

for

Human-Soybean Allergies: Elucidation of the Seed Proteome & Comprehensive Protein-Protein Interaction Prediction

Kevin Dick¹, Julia Hooker^{2,3¥}, Nour Nissan^{2,3¥}, Arezo Pattang^{2,3¥}, Michael Sadowski^{2,3}, Bradley Barnes¹, Le Hoa Tan^{2,3}, Daniel Burnside³, Sadhna Phanse⁴, Hiroyuki Aoki⁴, Mohan Babu⁴, Frank Dehne⁵, Ashkan Golshani³, Elroy R. Cober², James R. Green¹, Bahram Samanfar^{2,3,*}

¹ *Department of Systems and Computer Engineering, Carleton University, Ottawa, Ontario, Canada, K1S 5B6*

² *Agriculture and Agri-Food Canada, Ottawa Research and Development Centre, Ottawa, Ontario, Canada, K1A*

0C6

³ *Department of Biology and Institute of Biochemistry, and Ottawa Institute of Systems Biology, Carleton*

University, Ottawa, Ontario, Canada, K1S 5B6

⁴ *Department of Biochemistry, University of Regina, Regina, Saskatchewan, Canada, S4S 0A2*

⁵ *School of Computer Science, Carleton University, Ottawa, Ontario, Canada, K1S 5B6*

¥ These authors contributed equally to the work

*To whom correspondence should be addressed:

Bahram Samanfar

Agriculture and Agri-Food Canada

Ottawa Research and Development Centre (Ottawa RDC)

960 Carling Avenue, Neaty Building (#20)

Ottawa, Canada , K1A 0C6

Email: Bahram.samanfar@agr.gc.ca

Keywords: Soybean, human allergies, seed proteome, mass spectrometry, protein-protein interactions, computational biology, PIPE4.

Contents

The Supplementary tables and figures below summarize the results of applying the binomial test with Bonferoni multiple testing correction to the set of proteins predicted to interact with each of the Human Allergy Response (HAR) network and the Soybean Allergen (SA) network. These tests are used to qualify the level of overrepresentation of the predicted sample of proteins with respect to the reference proteome.

Tables:

Table S1: Predicted *Glycine max* Biological Process Overrepresentation Test

Table S2: Predicted Glycine max Molecular Function Overrepresentation Test

Table S3: Predicted Glycine max Cellular Component Overrepresentation Test

Table S4: Predicted Homo sapiens Biological Process Overrepresentation Test

Table S5: Predicted Homo sapiens Molecular Function Overrepresentation Test

Table S6: Predicted Homo sapiens Cellular Component Overrepresentation Test

Table S7: Soy-Seed Intersected Interactome (Table S7); Human-Seed Intersected Interactome

Table S8: Human-Seed Intersected Interactome

Table S9: Literature findings of immunological approaches that validate the key proteins involvement in human health

Table S10: Known Allergens from other species related to soybean

Table S11: All-to-All Predictions between Legume Proteins and Human Targets

Figures:

Figure S1: The Predicted Interactome for Proteins known to be involved in the Human Allergy Response

Figure S2: The Predicted Interactome for known Soybean Allergens

Figure S3: The Predicted Interactome for Proteins believed to be among the Soybean Seed Proteome

Table S1 (ST1): Predicted *Glycine max* Biological Process Overrepresentation Test

PANTHER GO-Slim Biological Process	Number in Reference	Number Predicted	Number Expected	Over/Under	Fold Enrichment	<i>p</i> Value
locomotion	46	20	1.87	+	10.69	3.25E-12
cellular component morphogenesis	115	40	4.68	+	8.55	1.12E-21
glycolysis	83	28	3.38	+	8.29	1.26E-14
regulation of carbohydrate metabolic process	58	17	2.36	+	7.2	1.15E-07
regulation of gene expression, epigenetic	75	20	3.05	+	6.55	1.89E-08
endocytosis	350	75	14.24	+	5.27	8.23E-28
chromatin organization	346	74	14.08	+	5.26	2.23E-27
cytoskeleton organization	259	53	10.54	+	5.03	1.49E-18
	159	32	6.47	+	4.95	1.01E-10
protein methylation	77	15	3.13	+	4.79	1.99E-04
cell cycle	1029	193	41.86	+	4.61	1.05E-64
protein phosphorylation	254	44	10.33	+	4.26	9.43E-13
glycogen metabolic process	275	47	11.19	+	4.2	1.85E-13
transcription initiation from RNA polymerase II promoter	84	14	3.42	+	4.1	2.54E-03
meiosis	104	17	4.23	+	4.02	4.15E-04
receptor-mediated endocytosis	104	16	4.23	+	3.78	1.70E-03
intracellular signal transduction	902	132	36.7	+	3.6	1.06E-32
cell surface receptor signaling pathway	214	31	8.71	+	3.56	6.09E-07
protein folding	361	52	14.69	+	3.54	4.53E-12
mitosis	463	65	18.84	+	3.45	9.33E-15
cell communication	1507	191	61.31	+	3.12	1.64E-39
signal transduction	1412	178	57.44	+	3.1	2.17E-36
apoptotic process	231	29	9.4	+	3.09	3.72E-05
DNA replication	247	31	10.05	+	3.08	1.45E-05
exocytosis	280	35	11.39	+	3.07	2.43E-06
polysaccharide metabolic process	616	70	25.06	+	2.79	1.63E-11
death	257	29	10.46	+	2.77	3.02E-04
cell death	257	29	10.46	+	2.77	3.02E-04
mRNA processing	607	66	24.69	+	2.67	5.46E-10
protein complex assembly	332	34	13.51	+	2.52	3.36E-04
protein complex biogenesis	337	34	13.71	+	2.48	4.61E-04
vesicle-mediated transport	1021	102	41.54	+	2.46	1.48E-13
RNA splicing, via transesterification reactions	432	42	17.58	+	2.39	8.48E-05
organelle organization	1798	170	73.15	+	2.32	6.56E-21
response to abiotic stimulus	306	28	12.45	+	2.25	1.75E-02
mRNA splicing, via spliceosome	463	42	18.84	+	2.23	4.69E-04
nucleobase-containing compound transport	289	26	11.76	+	2.21	3.93E-02
cellular protein modification process	1408	125	57.28	+	2.18	4.48E-13

regulation of nucleobase-containing compound metabolic process	863	76	35.11	+	2.16	1.90E-07
cellular component organization	2667	233	108.5	+	2.15	1.09E-24
intracellular protein transport	1099	95	44.71	+	2.12	4.25E-09
cellular component organization or biogenesis	3235	279	131.61	+	2.12	3.54E-29
cellular component biogenesis	1532	130	62.33	+	2.09	3.13E-12
transcription from RNA polymerase II promoter	885	74	36	+	2.06	2.55E-06
DNA metabolic process	634	52	25.79	+	2.02	5.78E-04
protein transport	1160	95	47.19	+	2.01	6.68E-08
translation	555	43	22.58	+	1.9	1.39E-02
protein metabolic process	3198	247	130.11	+	1.9	3.35E-19
generation of precursor metabolites and energy	510	39	20.75	+	1.88	3.75E-02
transcription, DNA-dependent	1642	118	66.8	+	1.77	9.62E-07
carbohydrate metabolic process	1309	94	53.25	+	1.77	3.56E-05
nucleobase-containing compound metabolic process	5562	384	226.28	+	1.7	2.71E-22
RNA metabolic process	3349	224	136.25	+	1.64	1.01E-10
response to stress	1576	104	64.12	+	1.62	3.64E-04
developmental process	1019	66	41.46	+	1.59	4.18E-02
response to stimulus	3433	220	139.67	+	1.58	7.92E-09
primary metabolic process	10614	649	431.81	+	1.5	2.26E-26
cellular process	14090	832	573.23	+	1.45	1.62E-31
catabolic process	2508	144	102.03	+	1.41	6.10E-03
regulation of biological process	2592	148	105.45	+	1.4	6.17E-03
biosynthetic process	4287	233	174.41	+	1.34	1.03E-03
metabolic process	14094	745	573.39	+	1.3	5.78E-14
nitrogen compound metabolic process	5930	297	241.25	+	1.23	2.26E-02
Unclassified	35065	1178	1426.56	-	0.83	0.00E+00
lipid metabolic process	1011	14	41.13	-	0.34	1.42E-04
ion transport	997	9	40.56	-	0.22	3.63E-07
anion transport	814	6	33.12	-	0.18	1.41E-06
steroid metabolic process	416	3	16.92	-	0.18	7.49E-03
cell differentiation	342	2	13.91	-	0.14	1.77E-02
homeostatic process	405	1	16.48	-	0.06	2.09E-04

Table S2 (ST2): Predicted *Glycine max* Molecular Function Overrepresentation Test

PANTHER GO-Slim Molecular Function	Number in Reference	Number Predicted	Number Expected	Over/ Under	Fold Enrichment	<i>p</i> Value
deacetylase activity	28	10	1.14	+	8.78	5.48E-05
nucleotide binding	300	79	12.2	+	6.47	2.53E-35
translation elongation factor activity	63	16	2.56	+	6.24	2.23E-06
calcium ion binding	143	33	5.82	+	5.67	9.31E-13
structural constituent of cytoskeleton	308	62	12.53	+	4.95	1.50E-21
translation regulator activity	196	35	7.97	+	4.39	2.02E-10
DNA-directed RNA polymerase activity	118	19	4.8	+	3.96	1.15E-04
transcription cofactor activity	212	33	8.62	+	3.83	2.83E-08
translation initiation factor activity	110	17	4.48	+	3.8	7.31E-04
structural molecule activity	851	117	34.62	+	3.38	1.33E-26
phosphoprotein phosphatase activity	332	43	13.51	+	3.18	1.69E-08
structural constituent of ribosome	498	55	20.26	+	2.71	1.68E-08
GTPase activity	381	38	15.5	+	2.45	1.37E-04
chromatin binding	342	34	13.91	+	2.44	5.34E-04
small GTPase regulator activity	212	21	8.62	+	2.43	3.79E-02
phosphatase activity	513	43	20.87	+	2.06	2.02E-03
kinase activity	1617	108	65.78	+	1.64	1.16E-04
protein kinase activity	1154	76	46.95	+	1.62	7.52E-03
protein binding	3110	200	126.53	+	1.58	4.23E-08
binding	6820	434	277.46	+	1.56	5.24E-19

DNA binding	1917	114	77.99	+	1.46	8.82E-03
nucleic acid binding	3590	195	146.05	+	1.34	5.47E-03
transferase activity	3998	208	162.65	+	1.28	3.24E-02
catalytic activity	11708	549	476.32	+	1.15	1.82E-02
Unclassified	35150	1323	1430.02	-	0.93	0.00E+00
transporter activity	2299	42	93.53	-	0.45	1.50E-07
cation transmembrane transporter activity	689	9	28.03	-	0.32	4.03E-03
transmembrane transporter activity	1988	23	80.88	-	0.28	2.14E-12
RNA polymerase II transcription factor binding transcription factor activity	324	2	13.18	-	0.15	2.85E-02
receptor activity	799	1	32.51	-	0.03	3.14E-11

Table S3 (ST3): Predicted *Glycine max* Cellular Component Overrepresentation Test

PANTHER GO-Slim Cellular Component	Number in Reference	Number Predicted	Number Expected	Over/Under	Fold Enrichment	<i>p</i> Value
heterotrimeric G-protein complex	29	7	1.18	+	5.93	1.08E-02
vesicle coat	109	23	4.43	+	5.19	1.89E-08
nuclear chromosome	189	31	7.69	+	4.03	9.01E-09
cytoskeleton	605	76	24.61	+	3.09	2.23E-15
microtubule	284	35	11.55	+	3.03	9.06E-07
chromosome	357	39	14.52	+	2.69	3.31E-06
nucleoplasm	580	62	23.6	+	2.63	1.28E-09
ribosome	617	61	25.1	+	2.43	3.55E-08
cytosol	1575	135	64.08	+	2.11	1.22E-13
ribonucleoprotein complex	1200	98	48.82	+	2.01	1.08E-08
macromolecular complex	3792	284	154.27	+	1.84	4.35E-21
protein complex	2746	201	111.72	+	1.8	1.38E-13
nucleus	4124	281	167.78	+	1.67	1.93E-15
cytoplasm	7113	468	289.38	+	1.62	3.02E-24
intracellular	11624	665	472.9	+	1.41	3.47E-20
organelle	8898	494	362	+	1.36	1.00E-11
cell part	12179	665	495.48	+	1.34	1.92E-15
Unclassified	38841	1499	1580.18	-	0.95	0.00E+00
membrane	4952	124	201.46	-	0.62	3.03E-08
plasma membrane	3432	77	139.63	-	0.55	9.28E-08
vacuole	721	12	29.33	-	0.41	1.13E-02

Table S4 (ST4): Predicted *Homo sapiens* Biological Process Overrepresentation Test

PANTHER GO-Slim Biological Process	Number in Reference	Number Predicted	Number Expected	Over/Under	Fold Enrichment	<i>p</i> Value
transmembrane receptor protein serine/threonine kinase signaling pathway	67	4	0.24	+	16.53	2.57E-02
intracellular signal transduction	1071	15	3.87	+	3.88	1.37E-03
cell communication	2686	24	9.7	+	2.47	3.62E-03
signal transduction	2318	20	8.37	+	2.39	3.77E-02

Unclassified	10206	25	36.86	-	0.68	0.00E+00
--------------	-------	----	-------	---	------	----------

Table S5 (ST5): Predicted *Homo sapiens* Molecular Function Overrepresentation Test

PANTHER GO-Slim Molecular Function	Number in Reference	Number Predicted	Number Expected	Over/Under	Fold Enrichment	<i>p</i> Value
ligand-activated sequence-specific DNA binding RNA polymerase II transcription factor activity	38	4	0.14	+	29.14	2.18E-03
sequence-specific DNA binding transcription factor activity	805	12	2.91	+	4.13	5.50E-03
receptor binding	865	11	3.12	+	3.52	4.85E-02
Unclassified	11852	35	42.81	-	0.82	0.00E+00

Table S6 (ST6): Predicted *Homo sapiens* Cellular Component Overrepresentation Test

PANTHER GO-Slim Cellular Component	Number in Reference	Number Predicted	Number Expected	Over/Under	Fold Enrichment	<i>p</i> Value
neuromuscular junction	16	4	0.06	+	69.22	2.59E-05
Unclassified	13830	51	49.95	+	1.02	0.00E+00

Table S7 (ST7): Soy-Seed Intersected Interactome (n=151)

Human	Soybean	SW Score
P05067	Glyma13g36400.1	9.74260698
P05067	Glyma20g28650.2	8.98387618
P05067	Glyma20g28650.1	8.77508941
P05067	Glyma06g43640.1	8.15211019
P05067	Glyma10g39150.1	8.00691658
P05067	Glyma17g03360.1	7.79419287
P05067	Glyma13g36780.1	7.66229216
P05067	Glyma10g39170.1	7.63844785
P05067	Glyma13g18450.1	7.60280751
P05067	Glyma12g34160.1	7.57317694
P05067	Glyma10g39170.2	7.19327576

P05067	Glyma02g01590.1	7.11553518
P05067	Glyma20g28640.1	6.89335489
P05067	Glyma20g28460.1	6.84752711
P05067	Glyma20g28460.2	6.84339248
P05067	Glyma03g32030.1	6.83974868
P05067	Glyma10g04280.1	6.73484619
Q06481	Glyma13g36400.1	6.48655914
P05067	Glyma19g42490.1	6.45007301
P05067	Glyma07g37270.1	6.40563661
P01040	Glyma08g12270.1	6.38115331
P05067	Glyma03g39940.1	6.36070207
Q06481	Glyma20g28650.2	5.87853112
Q06481	Glyma20g28650.1	5.7510872
P05067	Glyma19g34780.1	5.44945556
Q06481	Glyma06g43640.1	5.32082678
P62993	Glyma12g34160.1	5.28576203
Q06481	Glyma10g39150.1	5.2703667
P84022	Glyma07g37270.1	5.19117908
P84022	Glyma10g39150.1	5.15988986
Q06481	Glyma13g36780.1	5.05272737
P63279	Glyma13g36400.1	5.05020444
Q06481	Glyma17g03360.1	5.04625977
P51693	Glyma13g36400.1	5.04072558
Q06481	Glyma13g18450.1	5.04059071
Q13588	Glyma12g34160.1	5.03034017
Q06481	Glyma10g39170.1	4.94765677
Q06481	Glyma12g34160.1	4.9406922
Q15796	Glyma07g37270.1	4.89767472
Q15796	Glyma10g39150.1	4.82327139
P49223	Glyma13g36400.1	4.80472765
Q06481	Glyma02g01590.1	4.75149062
P84022	Glyma06g43640.1	4.74229169
Q06481	Glyma10g39170.2	4.66514442
P51693	Glyma20g28650.2	4.63610827
P63279	Glyma13g18450.1	4.5974351
Q06481	Glyma10g04280.1	4.5741471
P49917	Glyma08g12270.1	4.57125374
Q06481	Glyma20g28640.1	4.54905274
Q06481	Glyma03g32030.1	4.54581752

P51693	Glyma20g28650.1	4.53259359
Q06481	Glyma20g28460.2	4.52403582
Q06481	Glyma20g28460.1	4.51886521
P84022	Glyma20g28650.2	4.51632774
P04637	Glyma20g28460.1	4.49199771
P04637	Glyma20g28460.2	4.47706128
P78362	Glyma06g43640.1	4.47355901
Q15796	Glyma06g43640.1	4.47346948
P04637	Glyma20g28640.1	4.44963076
P04080	Glyma08g12270.1	4.43122363
O15198	Glyma10g39150.1	4.42816356
O15198	Glyma07g37270.1	4.42298304
P49223	Glyma20g28650.2	4.37926796
O15198	Glyma06g43640.1	4.33893184
P49223	Glyma20g28650.1	4.2920039
Q15797	Glyma10g39150.1	4.28183335
P05067	Glyma08g12270.1	4.28115476
Q06481	Glyma03g39940.1	4.28018593
Q15796	Glyma20g28650.2	4.25553128
P84022	Glyma20g28650.1	4.24970578
Q15797	Glyma07g37270.1	4.24741104
Q99717	Glyma10g39150.1	4.2327034
Q06481	Glyma19g42490.1	4.22915693
P63279	Glyma06g43640.1	4.22704641
P51693	Glyma10g39150.1	4.20488493
Q99717	Glyma07g37270.1	4.20040649
Q06481	Glyma07g37270.1	4.1960625
P84022	Glyma13g18450.1	4.18152898
P51693	Glyma06g43640.1	4.1464381
Q15797	Glyma06g43640.1	4.14470086
Q99717	Glyma06g43640.1	4.09791986
Q8TC17	Glyma12g34160.1	4.06699346
P03372	Glyma10g39150.1	4.05689657
P63279	Glyma12g34160.1	4.0508358
P84022	Glyma03g32030.1	4.04467132
P27348	Glyma20g28650.2	4.02074485
P84022	Glyma10g39170.2	4.01035635
P51693	Glyma13g36780.1	4.00520135
Q15796	Glyma20g28650.1	4.00361485

P84022	Glyma10g39170.1	4.00033131
O00625	Glyma08g12270.1	3.9847581
P51693	Glyma13g18450.1	3.93636537
Q15796	Glyma13g18450.1	3.93522219
O15198	Glyma20g28650.2	3.91973675
P51693	Glyma17g03360.1	3.91836642
P84022	Glyma20g28460.1	3.91764602
P49223	Glyma06g43640.1	3.91472125
P49223	Glyma10g39150.1	3.88837209
P84022	Glyma17g03360.1	3.87452599
P51693	Glyma10g39170.1	3.86062017
P63279	Glyma13g36780.1	3.85450101
P63279	Glyma10g39150.1	3.83129377
P49223	Glyma02g01590.1	3.8301826
Q9UPE1	Glyma06g43640.1	3.80671399
P84022	Glyma20g28640.1	3.80655055
P51693	Glyma12g34160.1	3.79894192
Q15796	Glyma03g32030.1	3.78996755
P27348	Glyma20g28650.1	3.78551149
P49223	Glyma17g03360.1	3.78252826
Q15797	Glyma20g28650.2	3.78093356
Q15796	Glyma10g39170.2	3.75982926
P49223	Glyma13g18450.1	3.75520643
P49223	Glyma13g36780.1	3.74803114
Q15796	Glyma10g39170.1	3.74385272
Q99717	Glyma20g28650.2	3.73454287
P84022	Glyma03g39940.1	3.73094755
P61981	Glyma20g28650.2	3.7298145
P84022	Glyma19g42490.1	3.72493515
P63279	Glyma10g04280.1	3.72090563
P62993	Glyma20g28650.2	3.70641404
Q8IWF7	Glyma03g39940.1	3.70426296
P78362	Glyma10g39170.2	3.70178061
O15198	Glyma20g28650.1	3.70007009
Q15796	Glyma17g03360.1	3.69148253
Q12933	Glyma20g28650.2	3.68731703
P51693	Glyma02g01590.1	3.68471099
P84022	Glyma19g34780.1	3.67397053
Q15796	Glyma20g28460.1	3.67318771

P51693	Glyma10g04280.1	3.65946618
Q8IZH2	Glyma08g12270.1	3.65782948
P49223	Glyma10g39170.1	3.65046776
P51693	Glyma10g39170.2	3.64897143
P49223	Glyma03g32030.1	3.64144658
P78362	Glyma10g39170.1	3.63847507
Q96SB4	Glyma06g43640.1	3.63743241
O15198	Glyma13g18450.1	3.63594378
P63104	Glyma20g28650.2	3.63511227
P46109	Glyma12g34160.1	3.6346831
P78362	Glyma20g28650.2	3.62237237
Q06481	Glyma19g34780.1	3.61764826
P62993	Glyma20g28650.1	3.60569001
O15198	Glyma03g32030.1	3.60532338
P51693	Glyma03g32030.1	3.59909907
P49223	Glyma12g34160.1	3.59233193
P51693	Glyma20g28640.1	3.58848389
Q04917	Glyma20g28650.2	3.58337615
P78362	Glyma20g28650.1	3.58208472
P03372	Glyma20g28650.2	3.57446688
Q15797	Glyma20g28650.1	3.57322579
Q15796	Glyma20g28640.1	3.56438138
P51693	Glyma20g28460.1	3.56372727

Table S8 (ST8): Human-Seed Intersected Interactome (n=134)

Human	Soybean	SW Score
P02788	Glyma10g28890.2	16.4344529
P02788	Glyma10g28890.1	15.8493822
Q99650	Glyma19g22780.1	14.8847517
Q99650	Glyma18g01330.2	14.6850836
Q99650	Glyma11g37360.1	14.6548676
Q99650	Glyma18g01330.1	14.5940161
Q99650	Glyma06g18110.3	14.4872593
Q99650	Glyma04g36870.2	14.4748496
P04179	Glyma10g39050.1	14.3734682
Q99650	Glyma06g18110.1	14.3706799
Q99650	Glyma04g36870.1	14.359904

Q99650	Glyma06g18110.4	14.1817023
P04179	Glyma11g00890.1	14.1301144
Q99650	Glyma06g18110.6	14.0898247
Q99650	Glyma06g18110.5	14.0688287
P07737	Glyma19g22850.1	13.0152485
Q05513	Glyma04g10820.2	12.8904159
P19878	Glyma05g26400.1	12.8462577
P19878	Glyma08g09310.1	12.6289237
P40763	Glyma17g10710.1	12.3959215
Q99650	Glyma06g18110.2	12.3447035
P07737	Glyma10g35700.1	11.9528213
P07737	Glyma05g24110.1	11.6825133
P07737	Glyma19g07240.2	11.630648
P07737	Glyma19g07240.1	11.630648
Q05513	Glyma04g10820.3	11.5154024
P15918	Glyma10g29000.1	11.4214372
Q05513	Glyma04g10820.1	11.3410473
Q05513	Glyma04g10820.4	11.2588828
Q04206	Glyma09g34760.3	10.6942489
Q04206	Glyma09g34760.4	10.6942489
Q04206	Glyma09g34760.1	10.6942489
P24394	Glyma05g34070.1	10.5922128
Q04206	Glyma09g34760.2	10.229101
P19878	Glyma08g09310.2	10.1507199
P07737	Glyma09g17040.1	9.08504548
P07737	Glyma02g10170.1	9.07821229
P07737	Glyma15g05570.1	9.01073457
P40763	Glyma02g10470.1	8.36474536
P56537	Glyma05g34070.1	7.83511192
P07737	Glyma15g06100.1	7.66559201
P05155	Glyma19g37520.1	7.54805919
P05155	Glyma09g28100.1	7.44103638
Q8N138	Glyma19g22850.1	7.40540256
Q99650	Glyma04g01750.2	7.31803435
Q04206	Glyma07g00820.1	7.14470768
Q04206	Glyma15g01750.1	7.13338961
Q8N138	Glyma10g35700.1	7.040549
P07737	Glyma04g10820.2	6.98932507
P56537	Glyma08g39770.1	6.97241805

Q99650	Glyma04g01750.1	6.95410639
Q8N138	Glyma05g24110.1	6.90514716
Q8N138	Glyma19g07240.2	6.87714465
Q8N138	Glyma19g07240.1	6.87714465
O14543	Glyma04g10820.2	6.73806634
P02768	Glyma09g30410.1	6.71463316
P06239	Glyma02g42560.1	6.33965636
P07737	Glyma04g10820.3	6.1842783
P07737	Glyma04g10820.1	6.11220007
P40763	Glyma05g34070.1	6.11122425
O14543	Glyma04g10820.3	6.1067058
O14543	Glyma07g13710.1	5.93954444
O14543	Glyma04g10820.4	5.91554033
O14543	Glyma04g10820.1	5.90159526
P07737	Glyma04g10820.4	5.89407443
Q05513	Glyma18g01330.2	5.64163206
Q05513	Glyma19g22780.1	5.6150186
Q05513	Glyma18g01330.1	5.60247481
Q05513	Glyma11g37360.1	5.59196555
Q9Y6K9	Glyma19g22850.1	5.55838028
Q05513	Glyma06g18110.3	5.55155902
Q05513	Glyma04g36870.2	5.55005954
Q04206	Glyma19g00870.2	5.5492129
Q04206	Glyma19g00870.1	5.5492129
P00709	Glyma17g08020.1	5.50659757
Q05513	Glyma06g18110.1	5.49644668
Q05513	Glyma04g36870.1	5.49514462
P06239	Glyma19g37520.1	5.47877551
Q04206	Glyma07g35110.2	5.43674539
Q04206	Glyma07g35110.1	5.43674539
Q05513	Glyma06g18110.4	5.41784927
P06239	Glyma09g28100.1	5.41268564
P00709	Glyma02g36700.1	5.40670956
Q05513	Glyma06g18110.6	5.39703316
Q05513	Glyma06g18110.5	5.38516831
Q9Y6K9	Glyma10g35700.1	5.35875291
P00709	Glyma13g19330.1	5.34897152
Q9Y6K9	Glyma07g18500.1	5.34794667
P06239	Glyma05g34070.1	5.28452047

Q9Y6K9	Glyma19g07240.1	5.21803154
Q9Y6K9	Glyma19g07240.2	5.21803154
Q9Y6K9	Glyma05g24110.1	5.2180257
P00709	Glyma05g36600.1	4.96475289
Q05513	Glyma14g01530.1	4.93449988
Q05513	Glyma02g47210.1	4.92561561
Q05513	Glyma08g44590.1	4.91581203
P02788	Glyma10g37960.1	4.89720988
P02788	Glyma10g37960.2	4.89679245
Q05513	Glyma09g24410.1	4.8740658
Q05513	Glyma02g47210.2	4.7738931
Q05513	Glyma06g18110.2	4.75220992
Q05513	Glyma10g20880.1	4.65332882
P40763	Glyma14g01530.1	4.56132303
P40763	Glyma08g44590.1	4.5575703
P40763	Glyma02g47210.1	4.54937
P06733	Glyma09g30190.2	4.40194462
P20023	Glyma09g30190.2	4.35119439
P40763	Glyma09g24410.1	4.29515352
P40763	Glyma02g47210.2	4.24716317
P01563	Glyma02g38730.1	4.21837636
P06733	Glyma04g10820.2	4.18358434
P15918	Glyma17g31120.1	4.11241508
P40763	Glyma10g20880.1	4.05138396
Q04206	Glyma10g34490.1	4.03139135
P56537	Glyma09g17040.1	3.99181787
P01563	Glyma10g07710.1	3.96781344
P01563	Glyma14g36850.1	3.94400517
P56537	Glyma02g10170.1	3.94315766
P56537	Glyma15g05570.1	3.92435853
P47710	Glyma13g34650.1	3.87349398
Q8N138	Glyma15g04290.1	3.81037122
P06733	Glyma04g10820.1	3.79816366
Q8N138	Glyma15g04290.2	3.78618535
P06733	Glyma04g10820.3	3.7654357
P20023	Glyma09g30190.1	3.71424884
P20023	Glyma09g30190.4	3.71424884
P31146	Glyma10g29000.1	3.70358022
P06733	Glyma04g10820.4	3.64767305

Q04206	Glyma10g29000.1	3.63136643
P26038	Glyma04g10820.2	3.61604689
Q05513	Glyma17g37820.1	3.61512727
P80511	Glyma05g34790.1	3.6149298
P06733	Glyma10g40150.1	3.60239283
P06733	Glyma09g30190.3	3.5626159

Table S9 (ST9): Literature findings of Immunological approaches that validate the key proteins involvement In human health

Candidate gene	Function	Immunological approach supporting genes as an allergen
P05067	Amyloid-beta A4 protein	[1]
P51693	Amyloid-like protein 1	[2–4]
Q06481	Amyloid-like protein 2	[3,5–7]
Q15796	Mothers against decapentaplegic homolog 2 (MAD homolog 2)	[8]
P84022	Mothers against decapentaplegic homolog 3 (MAD homolog 3)	[8,9]
Glyma11g05670.2	Cyclin-dependent kinase inhibitor 1	[10–12]
Glyma18g48660.1	Protein of unknown function/ SUMO gene (Arabidopsis ortholog)	[13–15]

Table S10 (ST10): Known allergens from other species related to soybean

Entry	Protein names	Organism	Length
A0A059TB88	Fatty acid and retinol binding protein (Fragment)	Meloidogyne arenaria (Peanut root-knot nematode)	132
A0A059TBD8	Venom allergen-like protein (Fragment)	Meloidogyne arenaria (Peanut root-knot nematode)	55
A0A1S6WN66	Expansin-like B8	Arachis duranensis (Wild peanut)	250
A0A1S6WN73	Expansin-like B8	Arachis ipaensis (Wild peanut)	250
A5Z1Q8	Ara d 2.01 (Conglutin)	Arachis duranensis (Wild peanut)	160
A5Z1R0	Ara h 6 (Ara h 6 allergen)	Arachis hypogaea (Peanut)	145
B0YIU5	Ara h 8 allergen isoform	Arachis hypogaea (Peanut)	153
B1PYZ4	Ara h 8 allergen isoform 3	Arachis hypogaea (Peanut)	157
B2CMA3	Major allergen Ara h1 (Fragment)	Arachis hypogaea (Peanut)	22
B2ZGS2	Pathogenesis-related class 10 protein	Arachis hypogaea (Peanut)	157
B3EWP3	Defensin 1 (allergen Ara h 12) (Fragment)	Arachis hypogaea (Peanut)	71

B3EWP4	Defensin 2 (allergen Ara h 13.0101) (Fragment)	Arachis hypogaea (Peanut)	79
B3IXL2	Main allergen Ara h1 (Major allergen Ara h1)	Arachis hypogaea (Peanut)	614
B4XID4	Ara h 7 allergen	Arachis hypogaea (Peanut)	164
B5TYU1	Arachin Arah3 isoform	Arachis hypogaea (Peanut)	530
B6CEX8	Non-specific lipid-transfer protein	Arachis hypogaea (Peanut)	116
B6CG41	Non-specific lipid-transfer protein (Fragment)	Arachis hypogaea (Peanut)	92
B8LF85	Venom allergen-like protein-1	Meloidogyne arenaria (Peanut root-knot nematode)	294
C0HJZ1	Defensin 3 (allergen Ara h 13.0102)	Arachis hypogaea (Peanut)	72
D3K177	Profilin	Arachis hypogaea (Peanut)	131
E5G076	Ara h 1 allergen	Arachis hypogaea (Peanut)	619
E5G077	Ara h 3 allergen	Arachis hypogaea (Peanut)	512
L7QH52	Profilin	Arachis hypogaea (Peanut)	131
O82580	Glycinin (Fragment)	Arachis hypogaea (Peanut)	507
P43237	Allergen Ara h 1, clone P17 (Allergen Ara h I) (allergen Ara h 1)	Arachis hypogaea (Peanut)	614
P43238	Allergen Ara h 1, clone P41B (Allergen Ara h I) (allergen Ara h 1)	Arachis hypogaea (Peanut)	626
Q08169	Hyaluronidase (Hya) (EC 3.2.1.35) (Allergen Api m II) (Hyaluronoglucosaminidase) (allergen Api m 2)	Apis mellifera (Honeybee)	382
Q0GM57	Iso-Ara h3	Arachis hypogaea (Peanut)	512
Q0PKR4	Pathogenesis-related protein 10	Arachis hypogaea (Peanut)	157
Q2YHR1	Pathogenesis-related protein 10 (Fragment)	Arachis hypogaea (Peanut)	135
Q45W79	Calcium-binding pollen allergen	Arachis hypogaea (Peanut)	164
Q5XXQ4	PR10 protein	Arachis hypogaea (Peanut)	150
Q647G9	Conglutin (allergen Ara h 6)	Arachis hypogaea (Peanut)	145
Q6IWG5	Glycinin (Fragment)	Arachis hypogaea (Peanut)	510
Q6PSU2	Conglutin-7 (2S protein 1) (Seed storage protein SSP1) (Seed storage protein SSP2) (allergen Ara h 2)	Arachis hypogaea (Peanut)	172
Q6VT83	Ara h 8 allergen	Arachis hypogaea (Peanut)	157
Q8LKN1	Allergen Arah3/Arah4	Arachis hypogaea (Peanut)	538
Q9SQH1	Allergen	Arachis hypogaea (Peanut)	160
Q9SQI9	Profilin (allergen Ara h 5)	Arachis hypogaea (Peanut)	131

Table S11 (ST11): All-to-All Predictions between Legume Proteins and Human Targets

Legume Protein	Human Protein	PIPE-Score	SW-Score
A0A1S6WN73	Q9Y4H2	0.05069760	0.98210634
A0A1S6WN73	O00763	0.02014522	1.26891299
A0A1S6WN73	P14672	0.01850870	0.54963336
A0A1S6WN73	P13807	0.01258305	0.72163537
A0A1S6WN73	P11498	0.01201962	0.81093568
A0A1S6WN73	P11168	0.01770177	0.60083151
A0A1S6WN73	P04035	0.00714360	0.78751015
A0A1S6WN73	P01130	0.02416212	0.87846359
A0A1S6WN73	Q13131	0.05117043	0.89276896
A0A1S6WN73	Q13085	0.02120598	1.35932968
A0A1S6WN73	Q12778	0.03224972	0.73654333
A0A1S6WN73	Q12772	0.02492457	0.80781459
A0A1S6WN73	Q07869	0.02607044	0.96193561
A0A1S6WN73	P49841	0.05804752	0.91425117
A0A1S6WN73	P49327	0.02791270	2.25980975
A0A1S6WN73	P40763	0.06313660	1.95125115
A0A1S6WN73	P37231	0.02904708	1.09473928
A0A1S6WN73	P36956	0.02690660	0.83136840
A0A1S6WN73	P35575	0.00563539	0.63860498
A0A1S6WN73	P35558	0.00795446	0.71237607
A0A1S6WN73	P35568	0.02912786	0.93347209
B3EWP3	Q9Y4H2	0.04722400	1.27267160
B3EWP3	P14672	0.00494505	0.72625589
B3EWP3	P13807	0.00886544	0.72179666
B3EWP3	P11498	0.00846220	0.90847548
B3EWP3	P11168	0.00460777	0.77593298
B3EWP3	P04035	0.00420466	0.84444985
B3EWP3	P01130	0.00974115	0.50038873
B3EWP3	Q13131	0.07026353	1.22674501
B3EWP3	Q13085	0.00523123	0.72720737
B3EWP3	Q12778	0.03958031	0.96725327
B3EWP3	Q12772	0.01712258	0.84108049
B3EWP3	Q07869	0.01464794	0.87223745
B3EWP3	P49841	0.03731057	0.74597161
B3EWP3	P49327	0.00841153	3.09018860
B3EWP3	P40763	0.01661887	0.61110827
B3EWP3	P37231	0.01349319	0.77124881
B3EWP3	P36956	0.02252114	0.84744954

B3EWP3	P35575	0.00250341	0.58511607
B3EWP3	P35568	0.02152651	1.07895151
B3EWP3	P35558	0.00178594	0.55450313
B3EWP3	O00763	0.00667834	0.75013404
Q45W79	Q9Y4H2	0.30762072	0.67635879
Q45W79	Q13131	0.33717752	0.75551724
Q45W79	Q13085	0.13712194	0.65441667
Q45W79	Q12778	0.22358491	0.53632618
Q45W79	Q12772	0.19678530	0.96301555
Q45W79	Q07869	0.15557945	0.62990554
Q45W79	P49841	0.31154871	0.52602975
Q45W79	P49327	0.14203243	1.00068357
Q45W79	P40763	0.32880298	1.72227375
Q45W79	P37231	0.14762310	0.52195260
Q45W79	P36956	0.19806799	0.66587797
Q45W79	P35575	0.04523567	0.43385023
Q45W79	P35568	0.22879296	0.73165478
Q45W79	P35558	0.09949105	0.58111740
Q45W79	P14672	0.08516538	0.31103448
Q45W79	P13807	0.13982326	0.53985208
Q45W79	P11498	0.17294933	0.75572878
Q45W79	P11168	0.08493001	0.34563332
Q45W79	P04035	0.08475061	0.71404309
Q45W79	P01130	0.09684694	0.30663824
Q45W79	O00763	0.12771769	0.61761604
Q6IWG5	Q9Y4H2	0.17857292	1.77926869
Q6IWG5	Q13131	0.15433356	1.20775817
Q6IWG5	Q13085	0.09777105	1.85965514
Q6IWG5	Q12778	0.13219716	1.29623474
Q6IWG5	Q12772	0.09997785	1.26213555
Q6IWG5	Q07869	0.12615044	1.31950159
Q6IWG5	P14672	0.04764953	0.60543248
Q6IWG5	P13807	0.06214082	1.52073535
Q6IWG5	P11498	0.08739889	1.93280604
Q6IWG5	P11168	0.04440725	0.65107782
Q6IWG5	P04035	0.04242065	1.30539586
Q6IWG5	P01130	0.08420051	0.96240776
Q6IWG5	P49841	0.16637632	1.33639933
Q6IWG5	P49327	0.13598873	3.93568666
Q6IWG5	P40763	0.16621965	2.05573831
Q6IWG5	P37231	0.12883340	1.42876300

Q6IWG5	P36956	0.12135460	1.25380790
Q6IWG5	P35575	0.03700334	1.15006206
Q6IWG5	P35568	0.11644099	1.60357073
Q6IWG5	P35558	0.05135558	1.14135365
Q6IWG5	O00763	0.09871496	1.80267947
A0A1S6WN66	Q9Y4H2	0.05250928	0.96590950
A0A1S6WN66	P35575	0.00546889	0.59336817
A0A1S6WN66	P35568	0.03027825	0.91421634
A0A1S6WN66	P35558	0.00823444	0.69523235
A0A1S6WN66	P14672	0.01946285	0.54351975
A0A1S6WN66	P13807	0.01374067	0.71526245
A0A1S6WN66	P11498	0.01266206	0.78593279
A0A1S6WN66	P11168	0.01806181	0.59153058
A0A1S6WN66	P04035	0.00784103	0.77172846
A0A1S6WN66	P01130	0.02190239	0.70953977
A0A1S6WN66	P49841	0.06041174	0.89864084
A0A1S6WN66	P49327	0.03221391	2.37685789
A0A1S6WN66	P40763	0.05621365	1.67536503
A0A1S6WN66	P37231	0.02567117	0.97389236
A0A1S6WN66	P36956	0.02801188	0.83022474
A0A1S6WN66	Q13131	0.05414462	0.88893699
A0A1S6WN66	Q13085	0.02194826	1.25667814
A0A1S6WN66	Q12778	0.03320265	0.73262272
A0A1S6WN66	Q12772	0.02572324	0.80120533
A0A1S6WN66	Q07869	0.02245490	0.79622827
A0A1S6WN66	O00763	0.02088359	1.18056332
Q5XXQ4	Q9Y4H2	0.16176377	2.91001742
Q5XXQ4	P49841	0.15021606	2.91174735
Q5XXQ4	P49327	0.12743374	6.23209844
Q5XXQ4	P40763	0.10451205	3.85173967
Q5XXQ4	P37231	0.06446141	1.98520403
Q5XXQ4	P36956	0.08954577	2.02143225
Q5XXQ4	P14672	0.04550553	1.41018850
Q5XXQ4	P13807	0.06188735	2.82213103
Q5XXQ4	P11498	0.07167932	3.64827536
Q5XXQ4	P11168	0.04515154	1.61776132
Q5XXQ4	P04035	0.03998630	2.60132292
Q5XXQ4	P01130	0.06060579	1.49150865
Q5XXQ4	P35575	0.01982926	1.43348841
Q5XXQ4	P35568	0.10921086	3.20604445
Q5XXQ4	P35558	0.03450939	1.80343828

Q5XXQ4	Q13131	0.11996042	2.58006785
Q5XXQ4	Q13085	0.06608778	3.83283853
Q5XXQ4	Q12778	0.10172116	2.05181478
Q5XXQ4	Q12772	0.07541059	2.14997074
Q5XXQ4	Q07869	0.07186453	1.90265050
Q5XXQ4	O00763	0.06966314	3.84174781
Q9SQI9	Q9Y4H2	0.11780434	0.91076844
Q9SQI9	P49841	0.10342447	0.93066441
Q9SQI9	P49327	0.03655268	2.09108074
Q9SQI9	P40763	0.03709340	0.69004423
Q9SQI9	P37231	0.03294018	0.82809744
Q9SQI9	P36956	0.06827033	0.92760575
Q9SQI9	P14672	0.04983601	1.06031341
Q9SQI9	P13807	0.02902407	0.78363758
Q9SQI9	P11498	0.06987243	2.25579317
Q9SQI9	P11168	0.04078854	1.08148868
Q9SQI9	P04035	0.01165132	0.79723204
Q9SQI9	P01130	0.08326609	1.26689103
Q9SQI9	P35575	0.01159658	0.65088757
Q9SQI9	P35568	0.06532531	0.86701320
Q9SQI9	P35558	0.03322672	1.52078891
Q9SQI9	Q13131	0.11319444	0.88705357
Q9SQI9	Q13085	0.03674259	1.08591688
Q9SQI9	Q12778	0.07867251	0.88291779
Q9SQI9	Q12772	0.05353164	0.86056468
Q9SQI9	Q07869	0.03863745	0.96042794
Q9SQI9	O00763	0.04150193	1.10678776
B3EWP4	Q9Y4H2	0.07413445	5.72011625
B3EWP4	P49841	0.03059019	0.91246883
B3EWP4	P49327	0.00693553	1.09258293
B3EWP4	P40763	0.01624501	0.80350644
B3EWP4	P37231	0.04427298	2.19838820
B3EWP4	P36956	0.02170508	1.14487293
B3EWP4	P14672	0.00901361	0.65897959
B3EWP4	P13807	0.00805478	0.69317549
B3EWP4	P11498	0.00670118	0.99965487
B3EWP4	P11168	0.00811881	0.68607261
B3EWP4	P04035	0.00617568	1.14919448
B3EWP4	P01130	0.04383670	3.53143084
B3EWP4	P35575	0.00350099	0.57800789
B3EWP4	P35568	0.03466885	3.06009812

B3EWP4	P35558	0.00489221	0.73095633
B3EWP4	Q13131	0.04447531	1.39361111
B3EWP4	Q13085	0.00677553	1.22593468
B3EWP4	Q12778	0.02604822	1.07649371
B3EWP4	Q12772	0.01356209	1.04273619
B3EWP4	Q07869	0.05883445	3.15267261
B3EWP4	O00763	0.00714090	1.12422441
B2ZGS2	Q9Y4H2	0.08014416	1.44735252
B2ZGS2	Q13131	0.08631240	1.55036232
B2ZGS2	Q13085	0.01516539	1.57095034
B2ZGS2	Q12778	0.05321985	1.16470695
B2ZGS2	Q12772	0.03461727	1.13818492
B2ZGS2	Q07869	0.04291340	1.03019593
B2ZGS2	P14672	0.01446318	0.88952972
B2ZGS2	P13807	0.02198135	1.07154536
B2ZGS2	P11498	0.03433120	1.93058109
B2ZGS2	P11168	0.01328742	0.88292438
B2ZGS2	P04035	0.03639032	1.42827004
B2ZGS2	P01130	0.02365197	0.59727895
B2ZGS2	P35575	0.00707486	0.84105137
B2ZGS2	P35568	0.04435517	1.41502838
B2ZGS2	P35558	0.01163266	0.93384527
B2ZGS2	P49841	0.08354115	1.31085330
B2ZGS2	P49327	0.08491521	4.00304162
B2ZGS2	P40763	0.09087400	2.25868890
B2ZGS2	P37231	0.03721596	1.01541719
B2ZGS2	P36956	0.04585518	1.12845616
B2ZGS2	O00763	0.01660220	1.49131861
B5TYU1	Q9Y4H2	0.21082953	1.16158241
B5TYU1	P14672	0.09381365	0.57510284
B5TYU1	P13807	0.09992150	0.91467383
B5TYU1	P11498	0.14137972	1.49005064
B5TYU1	P11168	0.08827188	0.59724865
B5TYU1	P04035	0.08481305	1.04358205
B5TYU1	P01130	0.12041624	0.76805173
B5TYU1	Q13131	0.19843444	1.07337102
B5TYU1	Q13085	0.14796606	1.78724276
B5TYU1	Q12778	0.17138672	1.06715775
B5TYU1	Q12772	0.14402747	1.13865197
B5TYU1	Q07869	0.15643374	1.10932753
B5TYU1	P35575	0.06869579	0.88575018

B5TYU1	P35558	0.08520671	0.90453473
B5TYU1	P35568	0.15302911	1.16319307
B5TYU1	P49841	0.21345365	1.20569418
B5TYU1	P49327	0.16955078	2.76997311
B5TYU1	P40763	0.20832497	1.94014764
B5TYU1	P37231	0.16255949	1.26772728
B5TYU1	P36956	0.16057376	1.10049999
B5TYU1	O00763	0.14703662	1.73498009
B1PYZ4	Q9Y4H2	0.11828790	1.69186142
B1PYZ4	P14672	0.03972198	1.20035492
B1PYZ4	P13807	0.03579791	1.33375722
B1PYZ4	P11498	0.07181353	3.31957209
B1PYZ4	P11168	0.03947482	1.31190989
B1PYZ4	P04035	0.04572139	1.96647821
B1PYZ4	P01130	0.03036413	0.76986507
B1PYZ4	Q13131	0.13138755	2.20261675
B1PYZ4	Q13085	0.04900569	2.60915342
B1PYZ4	Q12778	0.08123690	1.56329186
B1PYZ4	Q12772	0.05394094	1.43748870
B1PYZ4	Q07869	0.05485620	1.32545754
B1PYZ4	P35575	0.01796587	1.27420461
B1PYZ4	P35568	0.07125505	1.65646960
B1PYZ4	P35558	0.02227990	1.26634941
B1PYZ4	P49841	0.11404460	1.66050454
B1PYZ4	P49327	0.17452369	7.10610184
B1PYZ4	P40763	0.09003454	2.49584129
B1PYZ4	P37231	0.05609232	1.36213992
B1PYZ4	P36956	0.07018322	1.43609312
B1PYZ4	O00763	0.04974122	2.45308721
B4XID4	Q9Y4H2	0.28718203	1.17291051
B4XID4	Q13131	0.28463602	1.45137931
B4XID4	Q13085	0.17164323	1.29808693
B4XID4	Q12778	0.22516808	0.97359575
B4XID4	Q12772	0.16787141	0.90260004
B4XID4	Q07869	0.21033715	0.93481299
B4XID4	P35575	0.06376250	0.64054275
B4XID4	P35568	0.21099050	1.16636310
B4XID4	P35558	0.09482473	0.75634471
B4XID4	P49841	0.21872904	0.74109554
B4XID4	P49327	0.23195882	2.19955720
B4XID4	P40763	0.23038707	1.23030442

B4XID4	P37231	0.22189584	1.03492266
B4XID4	P36956	0.20353999	0.96182441
B4XID4	P14672	0.13835327	0.52836031
B4XID4	P13807	0.12464701	0.82807607
B4XID4	P11498	0.13782393	0.94114427
B4XID4	P11168	0.12129737	0.58307955
B4XID4	P04035	0.10200389	0.96634261
B4XID4	P01130	0.15757104	0.59697405
B4XID4	O00763	0.17066350	1.26314629
B8LF85	Q9Y4H2	0.08393135	0.92666345
B8LF85	Q13131	0.07564983	0.87779125
B8LF85	Q13085	0.02850021	1.00260343
B8LF85	Q12778	0.05805603	0.75144082
B8LF85	Q12772	0.03677200	0.79771836
B8LF85	Q07869	0.03728690	1.14154282
B8LF85	P35575	0.00730500	0.55217859
B8LF85	P35568	0.04481082	0.95084219
B8LF85	P35558	0.01221167	0.73345997
B8LF85	P49841	0.07214691	0.74899116
B8LF85	P49327	0.05442580	2.26377937
B8LF85	P40763	0.08577654	2.25946496
B8LF85	P37231	0.04077815	1.28995137
B8LF85	P36956	0.04575435	0.79897808
B8LF85	P14672	0.01130983	0.45708349
B8LF85	P13807	0.02019245	0.78411750
B8LF85	P11498	0.04986117	2.02530708
B8LF85	P11168	0.01193879	0.43663006
B8LF85	P04035	0.00883356	0.76683754
B8LF85	P01130	0.04571181	1.00589774
B8LF85	O00763	0.02910135	1.01340341
Q2YHR1	Q9Y4H2	0.08023973	0.69767457
Q2YHR1	Q13131	0.07289272	0.86130268
Q2YHR1	Q13085	0.01596328	0.71442808
Q2YHR1	Q12778	0.05434016	0.70738452
Q2YHR1	Q12772	0.03422921	0.63315201
Q2YHR1	Q07869	0.04863298	0.82434145
Q2YHR1	P49841	0.07859661	0.62505374
Q2YHR1	P49327	0.09750512	3.18925389
Q2YHR1	P40763	0.04420543	1.17442261
Q2YHR1	P37231	0.04109905	0.70602739
Q2YHR1	P36956	0.04408168	0.62363964

Q2YHR1	P35575	0.00813609	0.41261987
Q2YHR1	P35568	0.04310345	0.69805735
Q2YHR1	P35558	0.01242351	0.58161777
Q2YHR1	P14672	0.01377551	0.34062280
Q2YHR1	P13807	0.02077130	0.53381039
Q2YHR1	P11498	0.03327036	1.00145042
Q2YHR1	P11168	0.01300785	0.35988392
Q2YHR1	P04035	0.03974049	1.21209476
Q2YHR1	P01130	0.02589282	0.34816926
Q2YHR1	O00763	0.01812501	0.72466104
P43238	Q9Y4H2	0.25284119	1.75586817
P43238	P14672	0.11953737	0.76456645
P43238	P13807	0.12545374	1.19561247
P43238	P11498	0.15394598	2.08674040
P43238	P11168	0.11217969	0.83366663
P43238	P04035	0.11490797	1.53105029
P43238	P01130	0.15777091	1.29597620
P43238	P49841	0.25271253	1.62284569
P43238	P49327	0.20958732	3.30019621
P43238	P40763	0.20288380	1.97191224
P43238	P37231	0.17445983	1.67398187
P43238	P36956	0.18866329	1.44066272
P43238	Q13131	0.26988529	1.88233876
P43238	Q13085	0.17218612	2.30731708
P43238	Q12778	0.20921793	1.56238020
P43238	Q12772	0.17146217	1.65864674
P43238	Q07869	0.17785450	1.68222996
P43238	P35575	0.07416921	0.88771531
P43238	P35558	0.10435467	1.21417077
P43238	P35568	0.18191823	1.64679448
P43238	O00763	0.17126081	2.26480996
B0YIU5	Q9Y4H2	0.18436627	3.06051056
B0YIU5	P14672	0.04910143	1.43978069
B0YIU5	P13807	0.06935725	2.90176901
B0YIU5	P11498	0.13960182	4.91521255
B0YIU5	P11168	0.04792375	1.64492390
B0YIU5	P04035	0.04475036	2.69221785
B0YIU5	P01130	0.05588585	1.48830461
B0YIU5	P49841	0.20246027	4.10332378
B0YIU5	P49327	0.17085420	6.90815086
B0YIU5	P40763	0.17779279	4.88701632

B0YIU5	P37231	0.07530250	2.07597813
B0YIU5	P36956	0.09962819	2.08107733
B0YIU5	Q13131	0.13300166	2.61756495
B0YIU5	Q13085	0.10458665	4.27420162
B0YIU5	Q12778	0.11522576	2.13328405
B0YIU5	Q12772	0.08439088	2.23588608
B0YIU5	Q07869	0.08305355	1.97520194
B0YIU5	P35575	0.02578822	1.57985958
B0YIU5	P35568	0.12620056	3.35920357
B0YIU5	P35558	0.04050642	1.88831960
B0YIU5	O00763	0.11078372	4.32645200
Q0PKR4	Q9Y4H2	0.11740339	2.03463318
Q0PKR4	Q13131	0.10202630	1.37542941
Q0PKR4	Q13085	0.02506181	1.65785704
Q0PKR4	Q12778	0.06999134	1.23276137
Q0PKR4	Q12772	0.04580976	1.10919295
Q0PKR4	Q07869	0.03926600	1.04957877
Q0PKR4	P49841	0.13618129	1.85342802
Q0PKR4	P49327	0.09729395	4.07962291
Q0PKR4	P40763	0.07149887	1.69676181
Q0PKR4	P37231	0.03757381	0.97310193
Q0PKR4	P36956	0.06095179	1.17142050
Q0PKR4	P14672	0.01721384	0.59829932
Q0PKR4	P13807	0.02792580	0.98331719
Q0PKR4	P11498	0.05139363	2.78370909
Q0PKR4	P11168	0.01693213	0.66706845
Q0PKR4	P04035	0.01641901	1.25926852
Q0PKR4	P01130	0.02609040	0.58923125
Q0PKR4	P35575	0.01275620	0.99624818
Q0PKR4	P35568	0.09071302	3.48266321
Q0PKR4	P35558	0.01681207	1.03839498
Q0PKR4	O00763	0.02786245	1.64416695
L7QH52	Q9Y4H2	0.10457733	0.66573703
L7QH52	Q13131	0.10305886	0.73986442
L7QH52	Q13085	0.04373350	0.94513552
L7QH52	Q12778	0.07124607	0.70120171
L7QH52	Q12772	0.04906735	0.61881684
L7QH52	Q07869	0.02903277	0.54454343
L7QH52	P14672	0.02162901	0.53873907
L7QH52	P13807	0.03051632	0.64632660
L7QH52	P11498	0.04520523	1.54117620

L7QH52	P11168	0.01265912	0.47890736
L7QH52	P04035	0.01416858	0.59883076
L7QH52	P01130	0.07531425	1.56265925
L7QH52	P35575	0.01053994	0.44378698
L7QH52	P35568	0.06444195	0.67824290
L7QH52	P35558	0.01699834	0.79927742
L7QH52	P49841	0.08478803	0.53978892
L7QH52	P49327	0.02290186	1.52225694
L7QH52	P40763	0.03994674	0.58938083
L7QH52	P37231	0.02842078	0.48352072
L7QH52	P36956	0.06201716	0.68177717
L7QH52	O00763	0.04754217	0.94260675
A0A059TBD8	Q9Y4H2	0.03407464	1.00368545
A0A059TBD8	Q13131	0.02777778	0.70498971
A0A059TBD8	Q13085	0.00270974	0.69898057
A0A059TBD8	Q12778	0.02026555	0.73881901
A0A059TBD8	Q12772	0.01178451	0.73903248
A0A059TBD8	Q07869	0.01057906	0.93831972
A0A059TBD8	P14672	0.00090703	0.58407029
A0A059TBD8	P13807	0.00371402	1.36788146
A0A059TBD8	P11498	0.00337935	0.82892340
A0A059TBD8	P11168	0.00247525	0.64026403
A0A059TBD8	P04035	0.00086306	0.70614372
A0A059TBD8	P01130	0.00138724	0.35268860
A0A059TBD8	P35575	0.00139711	0.50041091
A0A059TBD8	P35568	0.01376397	1.22042791
A0A059TBD8	P35558	0.00138198	0.64602911
A0A059TBD8	P49841	0.02119701	0.59607925
A0A059TBD8	P49327	0.00276440	1.70955948
A0A059TBD8	P40763	0.00939488	1.22066874
A0A059TBD8	P37231	0.00903064	0.75377229
A0A059TBD8	P36956	0.01260835	0.76073680
A0A059TBD8	O00763	0.00404310	0.77909890
B2CMA3	Q9Y4H2	0.21101845	0.56456912
B2CMA3	Q13131	0.13209877	0.45185185
B2CMA3	Q13085	0.08895574	0.92837702
B2CMA3	Q12778	0.13784067	0.62316562
B2CMA3	Q12772	0.09774213	0.54961378
B2CMA3	Q07869	0.07943578	0.48700817
B2CMA3	O00763	0.08446084	0.81454148
B2CMA3	P14672	0.15646259	0.68503401

B2CMA3	P13807	0.06592386	0.53946147
B2CMA3	P11498	0.07736555	0.73482888
B2CMA3	P11168	0.09174917	0.70165017
B2CMA3	P04035	0.99846567	8.25661680
B2CMA3	P01130	0.55172414	4.53190646
B2CMA3	P49841	0.11305071	0.54945968
B2CMA3	P49327	0.08226324	0.88389513
B2CMA3	P40763	0.20594763	1.96671105
B2CMA3	P37231	0.11728395	0.82510288
B2CMA3	P36956	0.14716312	0.74083924
B2CMA3	P35575	0.01479290	0.62919132
B2CMA3	P35568	0.12183156	0.73889343
B2CMA3	P35558	0.04201216	0.66169154
P43237	Q9Y4H2	0.23850841	1.35721379
P43237	P14672	0.10187590	0.64620697
P43237	P13807	0.11418972	1.03655872
P43237	P11498	0.15272222	2.08336892
P43237	P11168	0.09405607	0.70141014
P43237	P04035	0.09882484	1.29903368
P43237	P01130	0.12797415	0.80749428
P43237	Q13131	0.26092094	1.72062290
P43237	Q13085	0.16608355	1.94480328
P43237	Q12778	0.18600841	1.16499905
P43237	Q12772	0.15189777	1.30697498
P43237	Q07869	0.14917925	1.09649577
P43237	P35575	0.06319606	0.75271452
P43237	P35558	0.09207889	1.06295961
P43237	P35568	0.16793858	1.27105891
P43237	P49841	0.23935532	1.40007725
P43237	P49327	0.18467243	2.64607147
P43237	P40763	0.20732850	1.78824418
P43237	P37231	0.14325699	1.03363886
P43237	P36956	0.17183806	1.17123958
P43237	O00763	0.16434883	1.91136664
E5G076	Q9Y4H2	0.24419383	1.59957796
E5G076	P14672	0.10409184	0.74020408
E5G076	P13807	0.11349350	1.18275998
E5G076	P11498	0.14342537	1.86136037
E5G076	P11168	0.09655116	0.83446205
E5G076	P04035	0.09380898	1.40998466
E5G076	P01130	0.13884463	1.00483948

E5G076	Q13131	0.27468519	1.90177160
E5G076	Q13085	0.16496490	2.12971709
E5G076	Q12778	0.19991876	1.39290618
E5G076	Q12772	0.15814023	1.45463161
E5G076	Q07869	0.16425019	1.45417595
E5G076	P35575	0.06305720	0.91445266
E5G076	P35558	0.08825594	1.18970149
E5G076	P35568	0.17158763	1.49911829
E5G076	P49841	0.25694514	1.61738155
E5G076	P49327	0.19950508	3.60244181
E5G076	P40763	0.19377941	1.84016423
E5G076	P37231	0.16445816	1.60479424
E5G076	P36956	0.17879728	1.33751625
E5G076	O00763	0.16268279	2.08558016
Q6PSU2	Q9Y4H2	0.23035375	1.34596421
Q6PSU2	Q13131	0.22829823	1.63833212
Q6PSU2	Q13085	0.15876988	2.15427027
Q6PSU2	Q12778	0.18446582	1.13763514
Q6PSU2	Q12772	0.14965107	1.20394836
Q6PSU2	Q07869	0.14804140	0.98439524
Q6PSU2	P35575	0.04140078	0.80359284
Q6PSU2	P35568	0.14870216	1.30235305
Q6PSU2	P35558	0.07109334	1.12808506
Q6PSU2	P14672	0.08047219	0.52411631
Q6PSU2	P13807	0.10144373	1.71017896
Q6PSU2	P11498	0.14992077	2.92736019
Q6PSU2	P11168	0.07681356	0.58949071
Q6PSU2	P04035	0.06061358	1.10899765
Q6PSU2	P01130	0.07704025	0.63896078
Q6PSU2	P49841	0.21368148	1.57679331
Q6PSU2	P49327	0.13838007	2.64527534
Q6PSU2	P40763	0.17921203	1.64084489
Q6PSU2	P37231	0.14123564	1.01507571
Q6PSU2	P36956	0.17647638	1.22736175
Q6PSU2	O00763	0.15151929	2.02344527
C0HJZ1	Q9Y4H2	0.07728840	6.42919879
C0HJZ1	Q13131	0.03990217	1.53581412
C0HJZ1	Q13085	0.00615417	1.00879746
C0HJZ1	Q12778	0.02385190	1.14079744
C0HJZ1	Q12772	0.01136784	1.07493358
C0HJZ1	Q07869	0.00798420	2.04403916

C0HJZ1	P35575	0.00117227	0.55185888
C0HJZ1	P35558	0.00406771	0.78891705
C0HJZ1	P35568	0.03596168	3.38871319
C0HJZ1	P49841	0.02187926	0.87422952
C0HJZ1	P49327	0.00663254	1.16436749
C0HJZ1	P40763	0.01487325	0.87428083
C0HJZ1	P37231	0.01199627	1.56277661
C0HJZ1	P36956	0.01948682	1.20597484
C0HJZ1	P14672	0.00820177	0.71039661
C0HJZ1	P13807	0.00641194	0.74533558
C0HJZ1	P11498	0.00673971	1.06308301
C0HJZ1	P11168	0.00556697	0.69702970
C0HJZ1	P04035	0.00568860	1.22941138
C0HJZ1	P01130	0.04733807	4.02788684
C0HJZ1	O00763	0.00622742	0.90760209
Q8LKN1	Q9Y4H2	0.21536722	1.22654811
Q8LKN1	Q13131	0.19926497	0.97455220
Q8LKN1	Q13085	0.15691394	1.80315108
Q8LKN1	Q12778	0.17388604	1.06920057
Q8LKN1	Q12772	0.15209387	1.17877002
Q8LKN1	Q07869	0.16857843	1.20365102
Q8LKN1	P35575	0.07366807	0.89778933
Q8LKN1	P35568	0.15892882	1.27033874
Q8LKN1	P35558	0.10227284	0.92748525
Q8LKN1	P49841	0.21876426	1.14703607
Q8LKN1	P49327	0.16355845	2.60852686
Q8LKN1	P40763	0.18708774	1.70089976
Q8LKN1	P37231	0.17446895	1.30719491
Q8LKN1	P36956	0.16896753	1.16878134
Q8LKN1	P14672	0.09784122	0.58277299
Q8LKN1	P13807	0.10352564	0.97739922
Q8LKN1	P11498	0.13581903	1.42771242
Q8LKN1	P11168	0.09297774	0.63371297
Q8LKN1	P04035	0.09112860	1.05163067
Q8LKN1	P01130	0.11849826	0.74606338
Q8LKN1	O00763	0.15394350	1.74353493
Q0GM57	Q9Y4H2	0.17785464	1.77237350
Q0GM57	P14672	0.04745622	0.60358488
Q0GM57	P13807	0.06189438	1.51671027
Q0GM57	P11498	0.08704433	1.92567209
Q0GM57	P11168	0.04422710	0.64893861

Q0GM57	P04035	0.04224856	1.30084707
Q0GM57	P01130	0.08385892	0.95888455
Q0GM57	Q13131	0.15380512	1.20494704
Q0GM57	Q13085	0.09737441	1.85298694
Q0GM57	Q12778	0.13166724	1.29134295
Q0GM57	Q12772	0.09959034	1.25792106
Q0GM57	Q07869	0.12563867	1.31414864
Q0GM57	P49841	0.16572666	1.33231829
Q0GM57	P49327	0.13543705	3.92026574
Q0GM57	P40763	0.16555613	2.04767680
Q0GM57	P37231	0.12831075	1.42391005
Q0GM57	P36956	0.12087847	1.24924654
Q0GM57	P35575	0.03685322	1.14698081
Q0GM57	P35568	0.11598354	1.59777855
Q0GM57	P35558	0.05116069	1.13910165
Q0GM57	O00763	0.09831449	1.79598928
Q08169	Q9Y4H2	0.12239425	1.33044067
Q08169	P14672	0.04430764	0.69222466
Q08169	P13807	0.03231735	0.77245870
Q08169	P11498	0.05717145	2.02738420
Q08169	P11168	0.04238606	0.76238169
Q08169	P04035	0.01600903	0.93105973
Q08169	P01130	0.07055748	1.18864463
Q08169	Q13131	0.12395164	1.81415672
Q08169	Q13085	0.05250379	1.97152128
Q08169	Q12778	0.08148812	0.92962212
Q08169	Q12772	0.05738719	0.95999863
Q08169	Q07869	0.05337235	0.91631234
Q08169	P49841	0.09208384	0.86078880
Q08169	P49327	0.10110701	3.60329694
Q08169	P40763	0.07104210	1.65430115
Q08169	P37231	0.04747815	0.86205489
Q08169	P36956	0.07362064	1.02237804
Q08169	P35575	0.01309762	0.65189822
Q08169	P35558	0.02020202	0.82684374
Q08169	P35568	0.07162534	1.42365677
Q08169	O00763	0.05334119	1.92191511
A0A059TB88	Q9Y4H2	0.26778801	1.59723443
A0A059TB88	Q13131	0.21086529	1.15696493
A0A059TB88	Q13085	0.07606740	1.15464478
A0A059TB88	Q12778	0.15534035	1.02885846

A0A059TB88	Q12772	0.10993327	0.96252741
A0A059TB88	Q07869	0.12474131	1.23895776
A0A059TB88	P14672	0.05378364	0.53640961
A0A059TB88	P13807	0.07132645	0.84996426
A0A059TB88	P11498	0.07051395	1.06000748
A0A059TB88	P11168	0.05164286	0.59365636
A0A059TB88	P04035	0.05397314	1.10386264
A0A059TB88	P01130	0.12345185	1.07279577
A0A059TB88	P35575	0.02434937	0.74977745
A0A059TB88	P35568	0.17437898	1.45571241
A0A059TB88	P35558	0.07762075	1.29271049
A0A059TB88	P49841	0.18438417	0.88733917
A0A059TB88	P49327	0.11037799	2.09585008
A0A059TB88	P40763	0.18622957	1.83608875
A0A059TB88	P37231	0.13778725	1.52197822
A0A059TB88	P36956	0.14655118	1.00305969
A0A059TB88	O00763	0.07564757	1.09145631
Q9SQH1	Q9Y4H2	0.24156491	1.12542814
Q9SQH1	Q13131	0.26677174	1.40458366
Q9SQH1	Q13085	0.14002140	1.60492461
Q9SQH1	Q12778	0.19309514	1.03067710
Q9SQH1	Q12772	0.16266545	1.17368301
Q9SQH1	Q07869	0.19125243	1.09831146
Q9SQH1	P14672	0.14952960	0.70270661
Q9SQH1	P13807	0.11433454	0.87796084
Q9SQH1	P11498	0.11796058	1.03254823
Q9SQH1	P11168	0.13784144	0.75393582
Q9SQH1	P04035	0.09458169	1.03962327
Q9SQH1	P01130	0.13573001	0.67773927
Q9SQH1	P49841	0.20346297	0.89158310
Q9SQH1	P49327	0.13852555	1.91420204
Q9SQH1	P40763	0.18025139	1.50405606
Q9SQH1	P37231	0.21241573	1.28882468
Q9SQH1	P36956	0.18393818	1.13105478
Q9SQH1	P35575	0.07157245	0.71499014
Q9SQH1	P35558	0.10177246	0.90829540
Q9SQH1	P35568	0.17394733	1.09545763
Q9SQH1	O00763	0.13918331	1.54573000
O82580	Q9Y4H2	0.22203234	1.30504046
O82580	Q13131	0.21413175	1.28812614
O82580	Q13085	0.15127037	1.88014100

O82580	Q12778	0.17926139	1.19033857
O82580	Q12772	0.14633010	1.23782546
O82580	Q07869	0.16506627	1.24112326
O82580	P14672	0.09762881	0.66997742
O82580	P13807	0.10980239	1.12148957
O82580	P11498	0.13767345	1.65204953
O82580	P11168	0.09244035	0.72096251
O82580	P04035	0.09055302	1.28086268
O82580	P01130	0.12264137	0.88469279
O82580	P49841	0.21795409	1.26376681
O82580	P49327	0.18285892	3.33035550
O82580	P40763	0.22345616	2.30001801
O82580	P37231	0.17175589	1.39117841
O82580	P36956	0.16696823	1.26614093
O82580	P35575	0.06586478	0.96053812
O82580	P35568	0.15930157	1.31130786
O82580	P35558	0.09607359	1.34204320
O82580	O00763	0.15167379	1.84413291
D3K177	Q9Y4H2	0.10297303	0.62640122
D3K177	P49841	0.08403099	0.56107499
D3K177	P49327	0.02282662	1.42786918
D3K177	P40763	0.03925718	0.55745910
D3K177	P37231	0.02764918	0.45995003
D3K177	P36956	0.06069529	0.64933194
D3K177	Q13131	0.10173611	0.61109458
D3K177	Q13085	0.04405964	0.89249263
D3K177	Q12778	0.07074068	0.66763533
D3K177	Q12772	0.04872517	0.60186688
D3K177	Q07869	0.02851575	0.52927140
D3K177	P14672	0.02095481	0.49759475
D3K177	P13807	0.02906138	0.59892310
D3K177	P11498	0.04462745	1.49936059
D3K177	P11168	0.01290665	0.46276521
D3K177	P04035	0.01395282	0.58470327
D3K177	P01130	0.07394471	1.51870647
D3K177	P35575	0.01035503	0.42532227
D3K177	P35568	0.06391631	0.64160145
D3K177	P35558	0.01680585	0.79165186
D3K177	O00763	0.04753851	0.88649842
A5Z1R0	Q9Y4H2	0.20852738	1.40723492
A5Z1R0	Q13131	0.14447384	0.98201058

A5Z1R0	Q13085	0.16342999	2.46006849
A5Z1R0	Q12778	0.13630578	1.07323800
A5Z1R0	Q12772	0.09626376	1.02475030
A5Z1R0	Q07869	0.07803938	0.85085021
A5Z1R0	P14672	0.04195011	0.51844833
A5Z1R0	P13807	0.06887518	1.07419198
A5Z1R0	P11498	0.07000425	1.20150787
A5Z1R0	P11168	0.04122269	0.59866415
A5Z1R0	P04035	0.04149086	1.18776371
A5Z1R0	P01130	0.05950022	0.68790933
A5Z1R0	P49841	0.21856074	1.75303804
A5Z1R0	P49327	0.12480573	2.12844276
A5Z1R0	P40763	0.16506034	2.24898020
A5Z1R0	P37231	0.07789536	0.90814227
A5Z1R0	P36956	0.12723038	1.13550462
A5Z1R0	P35575	0.02310510	0.73281206
A5Z1R0	P35568	0.12989786	1.38912251
A5Z1R0	P35558	0.04981705	1.12382532
A5Z1R0	O00763	0.15544362	2.36508913
B3IXL2	Q9Y4H2	0.23792789	1.36134326
B3IXL2	P14672	0.10181101	0.64780998
B3IXL2	P13807	0.11395098	1.03936472
B3IXL2	P11498	0.14737567	1.89075340
B3IXL2	P11168	0.09422082	0.70726017
B3IXL2	P04035	0.09880960	1.30502558
B3IXL2	P01130	0.12633220	0.78601705
B3IXL2	P49841	0.24026069	1.41321486
B3IXL2	P49327	0.18775578	2.74709457
B3IXL2	P40763	0.20757533	1.79959494
B3IXL2	P37231	0.14403292	1.04316146
B3IXL2	P36956	0.17201263	1.17646463
B3IXL2	Q13131	0.25976657	1.72070339
B3IXL2	Q13085	0.16596404	1.92717279
B3IXL2	Q12772	0.15216825	1.31274884
B3IXL2	Q12778	0.18520427	1.16534274
B3IXL2	Q07869	0.14974079	1.10545938
B3IXL2	P35575	0.06332355	0.76067326
B3IXL2	P35558	0.09220285	1.06933122
B3IXL2	P35568	0.16769069	1.27686293
B3IXL2	O00763	0.16418631	1.89395296
E5G077	Q9Y4H2	0.18592363	1.89746212

E5G077	P14672	0.05414166	0.64007120
E5G077	P13807	0.06758971	1.56746258
E5G077	P11498	0.08855819	1.96405063
E5G077	P11168	0.05039664	0.68986404
E5G077	P04035	0.04752613	1.36983126
E5G077	P01130	0.09311816	1.08899866
E5G077	Q13131	0.15979641	1.30020284
E5G077	Q13085	0.10145213	2.06510834
E5G077	Q12778	0.13950017	1.40934402
E5G077	Q12772	0.10729898	1.36463068
E5G077	Q07869	0.12835375	1.41380666
E5G077	P49841	0.17015271	1.41307988
E5G077	P49327	0.14059921	3.93279102
E5G077	P40763	0.17407486	2.14864832
E5G077	P37231	0.13130327	1.51484570
E5G077	P36956	0.12873311	1.34411909
E5G077	P35575	0.04136611	1.17807890
E5G077	P35558	0.05569516	1.19110667
E5G077	P35568	0.12410542	1.73105903
E5G077	O00763	0.10150138	1.99442378
Q6VT83	Q9Y4H2	0.09309864	1.52202481
Q6VT83	Q13131	0.08542673	1.35959474
Q6VT83	Q13085	0.02178273	1.41584923
Q6VT83	Q12778	0.06178790	1.26379774
Q6VT83	Q12772	0.04102405	1.22141492
Q6VT83	Q07869	0.04888480	1.11076143
Q6VT83	P35575	0.01050510	0.96046651
Q6VT83	P35568	0.05145342	1.41200066
Q6VT83	P35558	0.01491336	1.02733915
Q6VT83	P14672	0.01659272	0.72548063
Q6VT83	P13807	0.02483751	1.04732348
Q6VT83	P11498	0.06734316	2.84041715
Q6VT83	P11168	0.01597073	0.71884058
Q6VT83	P04035	0.03757442	1.61133070
Q6VT83	P01130	0.02466008	0.56444192
Q6VT83	P49841	0.09924464	1.42654234
Q6VT83	P49327	0.10036174	4.40461651
Q6VT83	P40763	0.05801926	1.79502692
Q6VT83	P37231	0.04356772	1.08986402
Q6VT83	P36956	0.05321719	1.17160679
Q6VT83	O00763	0.02324545	1.37641050

Q647G9	Q9Y4H2	0.20852738	1.40723492
Q647G9	Q13131	0.14447384	0.98201058
Q647G9	Q13085	0.16342999	2.46006849
Q647G9	Q12778	0.13630578	1.07323800
Q647G9	Q12772	0.09626376	1.02475030
Q647G9	Q07869	0.07803938	0.85085021
Q647G9	P35575	0.02310510	0.73281206
Q647G9	P35568	0.12989786	1.38912251
Q647G9	P35558	0.04981705	1.12382532
Q647G9	P14672	0.04195011	0.51844833
Q647G9	P13807	0.06887518	1.07419198
Q647G9	P11498	0.07000425	1.20150787
Q647G9	P11168	0.04122269	0.59866415
Q647G9	P04035	0.04149086	1.18776371
Q647G9	P01130	0.05950022	0.68790933
Q647G9	P49841	0.21856074	1.75303804
Q647G9	P49327	0.12480573	2.12844276
Q647G9	P40763	0.16506034	2.24898020
Q647G9	P37231	0.07789536	0.90814227
Q647G9	P36956	0.12723038	1.13550462
Q647G9	O00763	0.15544362	2.36508913
A5Z1Q8	Q9Y4H2	0.22260040	1.29660338
A5Z1Q8	Q13131	0.21971369	1.59265826
A5Z1Q8	Q13085	0.16561061	2.22424087
A5Z1Q8	Q12778	0.17484054	1.09011330
A5Z1Q8	Q12772	0.15233056	1.18948559
A5Z1Q8	Q07869	0.14140170	0.94929631
A5Z1Q8	P14672	0.08759589	0.54190187
A5Z1Q8	P13807	0.10826962	1.83535826
A5Z1Q8	P11498	0.19019820	3.19030223
A5Z1Q8	P11168	0.07919388	0.58325960
A5Z1Q8	P04035	0.06230362	1.09941320
A5Z1Q8	P01130	0.09004815	0.69808823
A5Z1Q8	P35575	0.04425280	0.80702086
A5Z1Q8	P35568	0.14347350	1.21174533
A5Z1Q8	P35558	0.07846112	1.19912259
A5Z1Q8	P49841	0.21434004	1.65428627
A5Z1Q8	P49327	0.13110606	2.49523013
A5Z1Q8	P40763	0.18505822	1.61668130
A5Z1Q8	P37231	0.13891078	1.01342556
A5Z1Q8	P36956	0.17627383	1.23353956

A5Z1Q8	O00763	0.15692107	2.07740936
B6CG41	Q9Y4H2	0.08347960	1.53794386
B6CG41	P49841	0.05339391	0.93116524
B6CG41	P49327	0.06496955	3.22594494
B6CG41	P40763	0.03431042	1.21117049
B6CG41	P37231	0.01491065	0.63391397
B6CG41	P36956	0.03893423	1.00064364
B6CG41	Q13131	0.10563166	5.11283612
B6CG41	Q13085	0.01932643	1.10188319
B6CG41	Q12778	0.04581287	0.97430430
B6CG41	Q12772	0.02867922	0.87663907
B6CG41	Q07869	0.01540715	0.60460079
B6CG41	P35575	0.00474183	0.52691092
B6CG41	P35568	0.05159108	1.47419886
B6CG41	P35558	0.01231286	0.95522388
B6CG41	P14672	0.01627062	0.62541236
B6CG41	P13807	0.01753348	0.85274927
B6CG41	P11498	0.06637749	4.66956635
B6CG41	P11168	0.01510918	0.72947240
B6CG41	P04035	0.01150748	1.08362627
B6CG41	P01130	0.05143909	1.99280048
B6CG41	O00763	0.02077541	1.10028251
B6CEX8	Q9Y4H2	0.08958677	1.53196345
B6CEX8	Q13131	0.14167621	4.52416953
B6CEX8	Q13085	0.02231536	1.84510830
B6CEX8	Q12778	0.05362122	0.95437010
B6CEX8	Q12772	0.04090633	0.94928056
B6CEX8	Q07869	0.01914908	0.74082153
B6CEX8	P49841	0.06298686	0.73820603
B6CEX8	P49327	0.10691946	6.27719217
B6CEX8	P40763	0.09591335	2.81609401
B6CEX8	P37231	0.02248526	0.83110602
B6CEX8	P36956	0.05023031	1.05607041
B6CEX8	P14672	0.01744162	0.66494845
B6CEX8	P13807	0.02166671	1.28594607
B6CEX8	P11498	0.03408555	2.37457638
B6CEX8	P11168	0.01098295	0.62402776
B6CEX8	P04035	0.01169729	1.04319457
B6CEX8	P01130	0.01971144	1.13101732
B6CEX8	P35575	0.00460562	0.63792472
B6CEX8	P35568	0.05220389	1.48417361

B6CEX8	P35558	0.01350635	1.69345711
B6CEX8	O00763	0.02401271	1.80523537

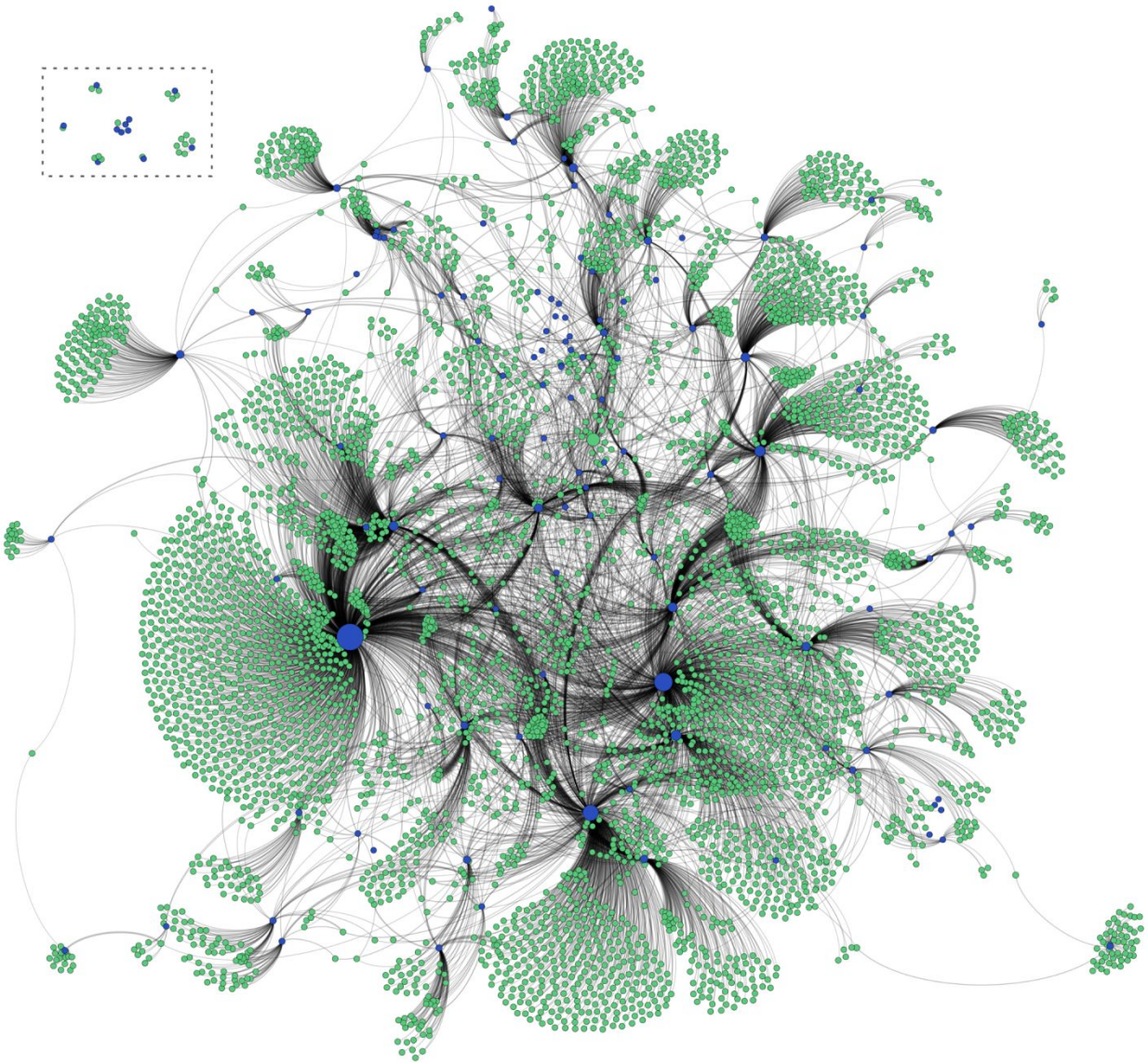


Figure S1 (SF1): The Predicted Interactome for Proteins known to be involved in the Human Allergy Response

Node size is proportional to the node betweenness centrality and edge weight is proportional to the predicted PPI score. The predicted network of 5,168 proteins and 8,054 edges comprises 97.52% soybean proteins and 2.48% human. The graph diameter (ignoring disjoint components) is 10. The box contains the disjoint networks from the

giant component. The interactive and query-able version of this network is available at <http://cu-bic.ca/human-soy/human>.

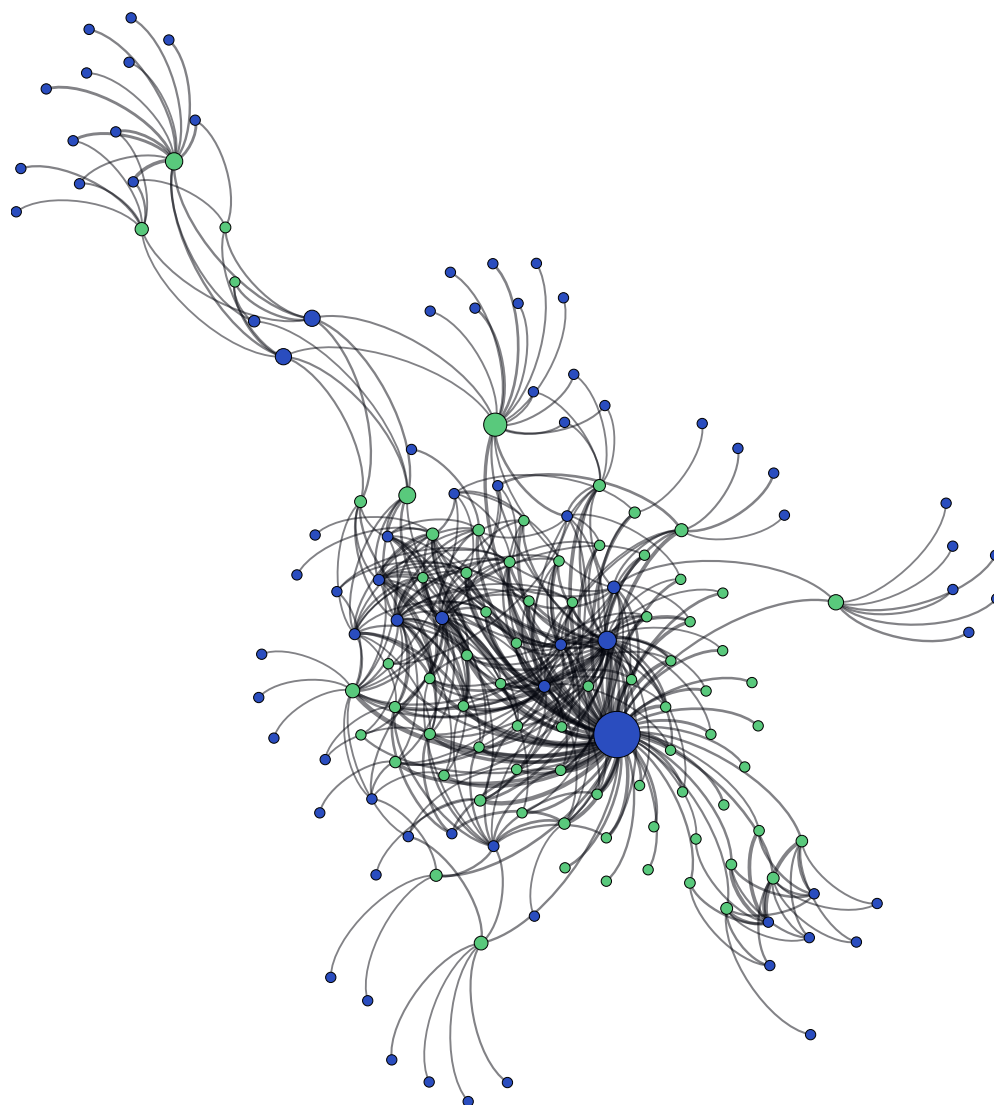


Figure S2 (SF2): The Predicted Interactome for known Soybean Allergens

Node size is proportional to the node betweenness centrality and edge weight is proportional to the predicted PPI score. The predicted network of 154 proteins and 438 edges comprises 49.35% soybean proteins and 50.65% human. The graph diameter is 6. The interactive and query-able version of this network is available at <http://cu-bic.ca/human-soy/soy>.

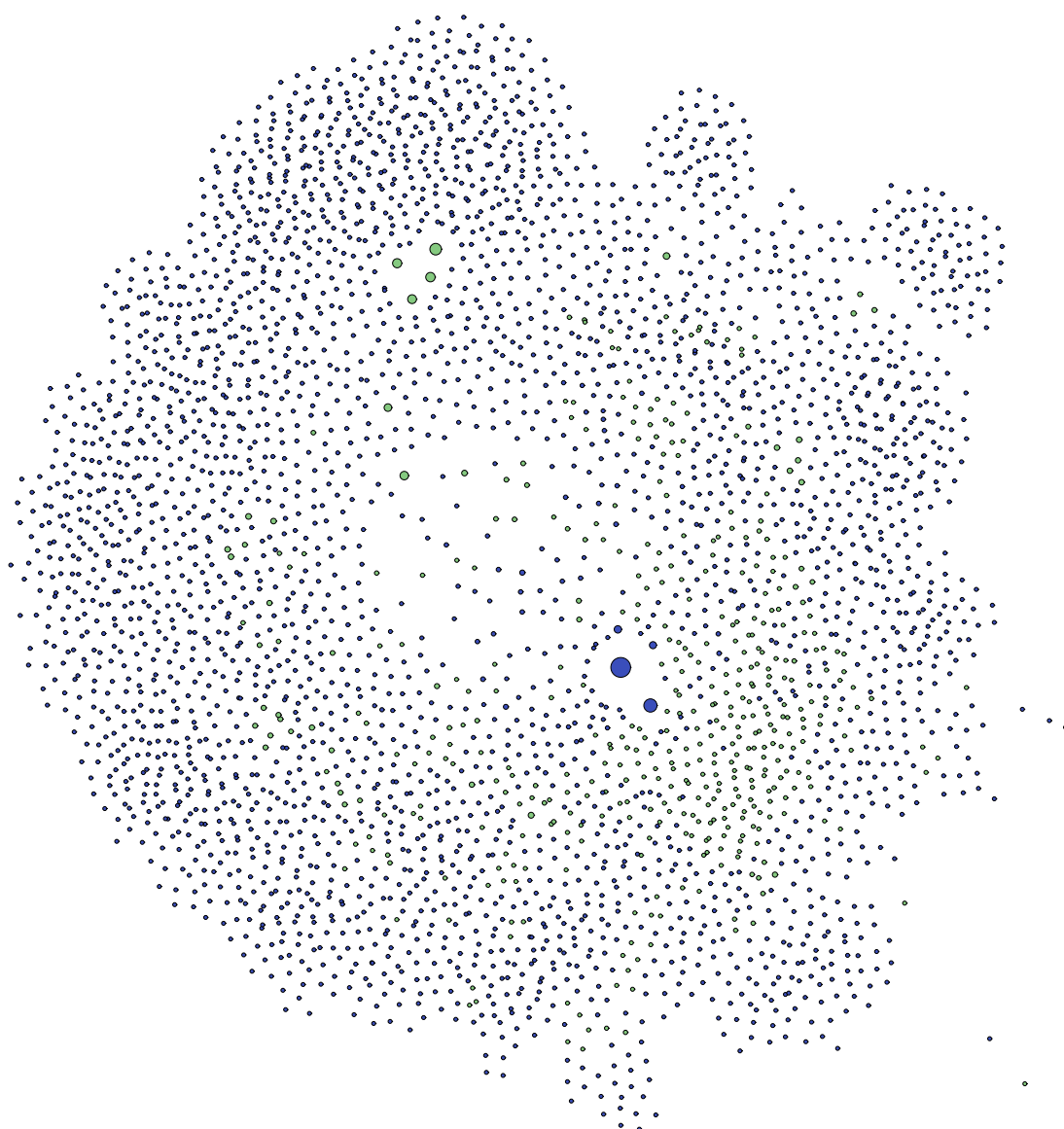


Figure S3 (SF3): The Predicted Interactome for Proteins believed to be among the Soybean Seed Proteome

Node size is proportional to the node betweenness centrality and edge weight is proportional to the predicted PPI score. The predicted network of 4,091 proteins and 19,351 edges comprises 11.17% soybean proteins and 88.83% human. The graph diameter (ignoring disjoint components) is 7. The disjoint networks are not depicted here, only the giant component. The interactive and query-able version of this network is available at <http://cu-bic.ca/human-soy/seed>.