



Assessment of the potential allergenicity and toxicity of *Pichia* proteins in a novel leghemoglobin preparation

Teresa F. Reyes^{*}, Yiming Chen, Rachel Z. Fraser, Teresa Chan, Xin Li

Impossible Foods Inc., Impossible Foods Inc., 400 Saginaw Drive, Redwood City, CA, 94063, USA

ARTICLE INFO

Keywords:

Leghemoglobin
Allergenicity
Pichia pastoris
Safety
Bioinformatics

ABSTRACT

The production of soy leghemoglobin C2 (LegH) by *Pichia pastoris* (syn. *K. phaffii*) was developed by Impossible Foods to serve as a sustainable source of flavor and aroma in plant-based meats. The potential allergenicity and toxicity of a LegH from a new production process was analyzed using bioinformatics, proteomics and a pepsin digestion assay on leghemoglobin, and residual host proteins. LegH in the new preparation had the same proteoform as in the previous preparations as well as in soy root nodule extracts. Results of seven *Pichia* proteins, each representing $\geq 1\%$ of the total protein content, showed no significant sequence matches to any known allergens with the exception of one, which matched the highly conserved wheat GAPDH, whose protein homolog is found in fungi and humans. Based on the data, it is unlikely that there is any risk of cross reactivity between LegH Prep and GAPDH. *Pichia* protein sequences showed very good alignment to homologous proteins from many common yeasts including *Saccharomyces* sp. In addition, LegH and *Pichia* proteins were all rapidly digested in a pepsin digest assay. In conclusion, LegH Prep from this *P. pastoris* production process is unlikely to pose a risk of food allergenicity.

1. Introduction

Leghemoglobins (LegHs) or legume hemoglobins are proteins naturally found in the root nodules of leguminous plants such as soy (Appleby, 1984, 1992). Each LegH protein consists of a cofactor heme bound to an iron, and one polypeptide chain (Hardison, 1998). Their main physiological function is to buffer oxygen levels as well as deliver oxygen to the nodules' bacteria dwellers to meet the high adenosine triphosphate (ATP) demands of nitrogen fixation (Hardison, 1998). This highly conserved mechanism is analogous to animal hemoglobin transporting oxygen to peripheral tissues (Krause et al., 1986; Hardison, 2012). Impossible Foods Inc. (IF) (Redwood City, CA.), pioneered the sustainable use of LegH in plant-based meat, which requires significantly less water, land and as a result, lower greenhouse gas emissions compared to animal-derived meat (Goldstein et al., 2017). IF's discovery that LegH is critical in imparting the meaty flavor and aroma in animal meat revolutionized the sensory experience of eating plant-based meat (Fraser et al., 2017). In addition to its role in flavor, heme is essential for human nutrition. About 95% of the functional iron in the human body,

as well as two-thirds of the average person's iron intake in developed countries is derived from heme (Hooda et al., 2014). To enable industrial-scale production of soy LegH, IF developed a yeast (*Pichia pastoris* syn. *Komagataella phaffii*) expression system. Soy LegH is delivered in a preparation referred to as "LegH Prep" which is a mixture containing the soy LegH C2 (Lbc2) isoform, residual *Pichia* proteins, and added food-grade stabilizers. The total protein fraction of LegH Prep contains at least 65% LegH with the remaining proteins belonging to the *Pichia* host (Fraser et al., 2018). The methylotrophic yeast *P. pastoris* is a non-toxicogenic, non-pathogenic industrial production organism with a long history of safe use dating back to the 1980's (Cereghino and Cregg, 2000; Lin Cereghino et al., 2001; Ciofalo et al., 2006). *Pichia* belongs to the family of yeast (*Saccharomycetaceae*) which is part of several yeast genera widely used in food such as *Saccharomyces*, *Torula*, and *Brettanomyces* (Kreger-van Rij, 1984), the latter which is commonly used in brewing and is part of the same sub-family of yeast as *Pichia* (*Pichiaceae*) (Colomer et al., 2020). Yeast extract mainly from *S. cerevisiae* and *Torula* species, is often directly consumed in substantial quantities in human diets. In 2006 a phospholipase C enzyme preparation from *P. pastoris*

^{*} Corresponding author

E-mail address: terry.reyes@impossiblefoods.com (T.F. Reyes).

<https://doi.org/10.1016/j.yrtph.2020.104817>

Received 17 June 2020; Received in revised form 16 October 2020; Accepted 30 October 2020

Available online 7 November 2020

0273-2300/© 2020 The Authors.

Published by Elsevier Inc.

This is an open access article under the CC BY-NC-ND license

(<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

received “a no questions” letter from FDA for use as food enzyme in degumming vegetable oils (U.S. Food and Drug Administration, 2006) and in 2010, the FDA granted approval of Kalbitor®, a pharmaceutical product used for treatment of acute attacks of hereditary angioedema (Thompson, 2010). The American Association of Feed Control Officials (AAFCO) has approved the enzyme phytase (from *E. coli*) derived from the fermentation of recombinant *P. pastoris* for use in animal feed (AAFCO, 2013). *P. pastoris* is also used for production of nitrate reductase (The Nitrate Elimination Co. Lake Linden, MI), an enzyme used for the treatment of potable water. In 1993, the US Food and Drug Administration (FDA) approved a food additive petition (FAP), submitted by the Phillips Petroleum Company, for use of dried *P. pastoris* yeast in broiler chicken feed formulations at up to 10% (U.S. Food Additive Petition (FAP) #2207, 1993). Supporting safety data included a short-term swine feeding study, which was conducted to evaluate the *P. pastoris* single cell protein (SCP) as a dietary protein source for pigs (Supplemental Table 1). In this paper, a summary of the results of this feeding study, which was obtained through the Freedom of Information Act (FOIA), is presented with the aim of sharing as well as to provide additional evidence that supports the safe consumption of *Pichia* proteins.

The safety of LegH derived from the yeast *P. pastoris* has been previously evaluated through a battery of *in vivo* and *in vitro* tests (Fraser et al., 2018). LegH Prep’s systemic toxicity was evaluated in a 28-day dietary study in Sprague Dawley rats. There were no mortalities, no clinical observations, body weight, ophthalmological, clinical pathology, or histopathological changes attributable to LegH Prep administration (Fraser et al., 2018). Additionally, genotoxicity tests were performed (Ames test and an *in vitro* chromosomal aberration) on LegH Prep. Results showed that LegH Prep was non-mutagenic and non-clastogenic. Overall, results of these studies have shown no issues of toxicological concern with regard to LegH Prep under the conditions tested (Fraser et al., 2018). The potential risk of food allergy and toxicity of soy LegH expressed in *Pichia* sp. has been previously evaluated (Jin et al., 2018). In this report, the strain MX0291, subject of GRN 737 (U.S. Food and Drug Administration, 2018) was assessed for potential allergenicity and toxicity via bioinformatics sequence comparisons to known allergens or toxins, an *in vitro* pepsin digestion assay and a literature review. Bioinformatics analyses resulted in 13 *Pichia* proteins that significantly matched to known allergens or toxins. However, all the matched allergens and toxins showed similar matches to proteins from the common yeast *S. cerevisiae*, without evidence of food allergy or toxicity. Results from the literature searches found no evidence of allergenicity or toxicity; and there were also no significant matches of LegH to known allergens or toxins (Jin et al., 2018). In 2018, FDA issued a “no questions letter” on the Generally Recognized as Safe (GRAS) food status of *P. pastoris* derived LegH Prep at use levels up to 0.8% soy leghemoglobin protein (U.S. Food and Drug Administration, 2018). In 2019, the FDA approved the use of soy LegH as a color additive (Federal Register and the Daily Journal of the United States Government, 2018).

This paper focuses on the safety, by means of allergenic potential of *Pichia*-derived proteins, from a new simplified manufacturing process to produce LegH Prep, which we will call LegH Prep 2.0. All raw materials used in the manufacture of LegH Prep 2.0 conform to Food Chemicals Codex (FCC) specifications, maintained by the U.S. Pharmacopeia (2018), or are food-grade ingredients. Various regulatory agencies, including the U.S. FDA, have updated their allergenicity assessment guidelines over the years; however, there isn’t a conclusive test for determining the allergenic potential of novel proteins; therefore, a “weight-of-evidence” approach, consistent with Codex Alimentarius “Guideline for the conduct of Food Safety Assessment of Foods derived from recombinant-DNA plants” is used (Alimentarius, 2003; Ladics, 2008). Proteomics analysis can serve as part of this approach since it is a

widely recognized tool for the identification of putative food protein allergens (Di Girolamo et al., 2015). To assess the potential allergenicity and toxicity of proteins in LegH Prep 2.0, we conducted a mass spectrometry-based analysis of the LegHs derived from both soy root nodule and *Pichia*, as well as the quantitative proteomics analysis on the LegH Prep 2.0 for the identification and quantification of the residual *Pichia* proteins, which were then investigated using bioinformatics methods. Additionally, an *in vitro* pepsin digestion assay was also performed. The allergenicity assessment was conducted following the most current allergenicity assessment guidelines developed by the Codex Commission (2009) (Alimentarius, 2009). Codex recommends a weight-of-evidence approach to determine whether an introduced new protein has a potential allergenicity risk (Ladics, 2008). This multi-step process involves the evaluation of the safety and allergenicity of the gene source, analysis of its sequence by comparison to known allergens and a digestion assay to test the stability of the protein when exposed to pepsin. Bioinformatics by sequence comparison in combination with a literature review can be used as powerful tools for evaluating the potential risks of food allergenicity (Alimentarius, 2009; Goodman, 2008). When a significant sequence identity is found, serum IgE tests using samples from source-sensitized subjects may be needed for evaluation of the allergenic potential (Goodman, 2008). Codex also recommends a review of toxins or toxicity of the source organism and a comparison of the protein amino acid sequence identity between the novel protein and known protein toxins (Mathesius et al., 2009). Similarly, Codex recommends that the stability of the protein be measured via an *in vitro* simulated gastric pepsin digestion assay to evaluate potential risks of allergenicity or toxicity of the protein (Alimentarius, 2009; Goodman, 2008; Mathesius et al., 2009). Collectively, the results of the studies presented show that LegH Prep 2.0, from the new production process, is unlikely to present any additional risks for allergenicity or toxicity.

2. Materials and methods

2.1. Materials

Impossible Foods Inc. provided three LegH Prep 2.0 samples from production runs: PP-PGM2-18-338, PP-PGM2-19-045, and PP-PGM2-19-056. All chemicals are technical grade and were purchased from Sigma-Aldrich unless noted otherwise, and used without further purification:

Urea (BioReagent, ≥98%), DL-dithiothreitol solution (1 M in H₂O), Supelco; Iodoacetamide (BioUltra, ≥99%), Acetone (CHROMASOLV Plus, ≥99.9%), Honeywell; Trypsin Gold (Mass Spectrometry Grade), Promega; Ammonium bicarbonate (BioUltra, ≥99.5%), Water (Optima LC/MS Grade), Fisher Chemical; Acetonitrile (Optima LC/MS Grade), Fisher Chemical; Formic acid (Optima LC/MS Grade, ≥99.0%), Fisher Chemical; Acetic acid (ACS reagent, ≥99.7%), and Ethyl acetate (anhydrous, 99.8%). Pepsin A (Worthington Chemicals) with a certificate of analysis of 2600 activity units/mg was used in all digestion assays. Precast polyacrylamide gel (AnykD™ Criterion™ TGX™ Protein Gels, Bio-Rad) were used to separate digested materials. Precision Plus Protein Dual Xtra Standards (Bio-Rad) were used as the molecular weight standards. Tris-Glycine-SDS 10x running buffer (Bio-Rad) and InstantBlue Protein Stain (Expedeon) were used for gel running and staining. A 3 kD Amicon Ultra-0.5 mL centrifugal filter (EMD Millipore) was used for protein buffer exchange.

2.2. Shotgun proteomics

2.2.1. Sample preparation for shotgun proteomics

Protein analysis was performed using shotgun proteomics (Zhang et al., 2013). Fifty micrograms of protein was added with 50 µL solution containing 8M urea and 10 mM DL-dithiothreitol. Proteins in the

mixture were dissolved and reduced for 45 min (min) at 50 °C, followed by alkylation with iodoacetamide (30 mM) for 30 min in the dark. Two hundred microliters of pre-chilled 100% acetone was added to precipitate. After being pelleted by 5 min centrifugation at the speed of 4000×g, the protein pellet was washed twice with pre-chilled 80% acetone and dried with a SpeedVac concentrator. Two micrograms of trypsin (Promega) in 20 µL of ammonium bicarbonate buffer (50 mM, pH = 7.8) was added and the digestion was carried out for 12 h at 37 °C. Digested protein was dried with a SpeedVac concentrator and solubilized with 25 µL of buffer (5% acetonitrile and 0.1% formic acid).

2.2.2. Liquid chromatography with tandem mass spectrometry (LC-MS/MS) data collection for shotgun proteomics

LC-MS/MS data were collected on a Q Exactive Plus mass spectrometer (Thermo Fisher Scientific) coupled with a UHPLC (Vanquish Ultra-High Performance Liquid Chromatography, Thermo Fisher Scientific). Five micrograms of peptides were injected into an ACQUITY UPLC CSH C18 column (1 × 100 mm, 1.7 µm, 130 Å, Waters) which was maintained at 40 °C and kept at 0.1 mL/min constant flow rate. Peptides were eluted with a gradient from buffer A (5% acetonitrile and 0.1% formic acid) to buffer B (0.1% formic acid in acetonitrile). The elution gradient program was: isocratic at 0% B for 1.5 min, 0–25% B in 28.5 min, 25–50% B in 9 min, 50–95% B in 3 min, isocratic at 95% B for 2 min, and isocratic at 0% B for 1.5 min. Q Exactive Plus Orbitrap employed a full MS scan followed by top-10 data-dependent MS² acquisition. Full MS scan parameters were 70,000 resolution, AGC (Automatic Gain Control) target of 3e6 counts, maximum fill times of 200 ms, and scan range of 400–2000 m/z. dd-MS². Scan parameters were 17,500 resolution, AGC target of 1e5 counts, maximum fill times of 50 ms, the normalized collision energy of 27, and dynamic exclusion of 30 s.

2.2.3. Data analysis of shotgun proteomics

The mass spectrometric raw data were analyzed with the MaxQuant software (Tyanova et al., 2016). The search database was built based on the high-quality genome sequence of *P. pastoris* CBS7435 (Küberl et al., 2011). The sequence of soy LegH C2 was manually added. The false discovery rate (FDR) was set to 0.01 for proteins and peptides. All identified proteins were quantified with LFQ (label free quantification) algorithm of MaxQuant.

2.3. Intact protein mass spectrometry (IPMS) analysis

2.3.1. Sample preparation

Fifty micrograms of LegH Prep 2.0 was first diluted by 50 mM ammonium bicarbonate buffer (pH = 7.8) 5 × , and then concentrated back to 50 µL in 3 kD Amicon Ultra-0.5 mL centrifugal filter (EMD Millipore). Then, the concentrate was diluted with 50 mM ammonium bicarbonate buffer (pH = 6.5) to five hundred microliters, then concentrated back to 50 µL again. The process was repeated three times to desalt the protein sample. Finally, the concentrated protein sample was diluted with 50 mM ammonium bicarbonate buffer (pH = 6.5) to five hundred microliters, and filtered using a 0.2 µm syringe filter, and then transferred to a plastic vial and kept frozen in –80 °C freezer upon analysis.

2.3.2. Liquid chromatography with full scan mass spectrometry (LC-MS) data collection

LC-MS data were collected on the same LC-MS platform as section 2.2.2. Five micrograms of intact desalted protein was injected into an ACQUITY UPLC Protein BEH C4 column (300 Å, 1.7 µm, 1 mm × 150 mm, Waters) which was maintained at 40 °C and kept at 0.1 mL/min constant flow rate. The elution gradient program was: isocratic at 0% B for 5 min, 0–80% B in 15 min, isocratic at 80% B for 5 min, and isocratic

at 0% B for 5 min. Q Exactive Plus Orbitrap employed a full MS scan only mode. Full MS scan parameters were 140,000 resolution with 5 micro scans, AGC target of 3e6 counts, maximum fill times of 150 ms, and scan range of 400–2600 m/z.

2.3.3. Data analysis of IPMS

The raw data acquired on the mass spectrometer was loaded into BioPharma Finder 3.0 (Thermo Fisher Scientific), sliding window deconvolution method was applied to the whole acquisition window (0–30 min) with the Xtract algorithm, with the merge tolerance window width of 30.0 ppm, and while the maximum retention time gap set to 1.00 min. Only the precursor ions detected in at least three intervals are counted for deconvolution. The output mass after deconvolution is limited in the range of 1000 Da to 60 kD, and the charge status of the precursor ions is restrained between 5+ and 50+. The minimum signal/noise ratio cutoff is set to three, and the minimum detected charges cutoff is set to three.

2.4. Top-down proteomics analysis

2.4.1. LegH extraction and purification from soy root nodule

Soybean root-nodule lysate was purified by 60/90% (wt./v) ammonium sulfate fractionation. Proteins from 90% ammonium sulfate pellet were further purified by anion-exchange chromatography (diethylaminoethyl, (DEAE) GE Healthcare) in 20 mM potassium phosphate pH 7.4, 100 mM sodium chloride. Anion-exchange flow through was fractionated using size-exclusion chromatography (Sephacryl S-100 GE Healthcare).

2.4.2. Sample preparation for the top-down proteomics

The sample preparation method is similar to the method described in 2.3.1. In general, 50 µg of purified LegH from soy root nodule was first diluted by 50 mM ammonium bicarbonate buffer (pH = 7.8) for 5 × , and then concentrated back to 50 µL in 3 kD Amicon Ultra-0.5 mL centrifugal filter (EMD Millipore). Then, the concentrate was diluted with 50 mM ammonium bicarbonate buffer (pH = 6.5) to five hundred microliters, and concentrated back to 50 µL again. The process was repeated three times to desalt the protein sample. Finally, the concentrated protein sample was diluted with 50 mM ammonium bicarbonate buffer (pH = 6.5) to five hundred microliters, filtered with 0.2 µm syringe filter, and transferred to a plastic vial and kept frozen in –80 °C freezer upon analysis.

2.4.3. Liquid chromatography with tandem mass spectrometry (LC-MS/MS) data acquisition

LC-MS data were collected on the same LC-MS platform as section 2.2.2. Fifteen micrograms of desalted intact protein was injected into an ACQUITY UPLC Protein BEH C4 column (300 Å, 1.7 µm, 1 mm × 150 mm, Waters) which was maintained at 40 °C and kept at 0.1 mL/min constant flow rate. The elution gradient program was isocratic at 0% B for 5 min, 0–95% B in 45 min, isocratic at 95% B for 10 min, and isocratic at 0% B for 5 min. Q Exactive Plus Orbitrap employed a full MS scan followed by top-2 data-dependent MS² acquisition. The full MS scan parameters were 70,000 resolution with 4 micro scans, AGC (Automatic Gain Control) target of 1e6 counts, maximum fill times of 50 ms, and a scan range of 400–2600 m/z. The dd-MS² scan parameters were: 35,000 resolution with 4 micro scans, AGC target of 1e6 counts, maximum fill times of 800 ms, isolation windows 3.0 m/z, the normalized collision energy of 22, and the dynamic exclusion of 100 s. The charge stats below 6+ were excluded from the MS/MS scan.

2.4.4. Data analysis of top-down proteomics

The raw data files were first converted to mzML format with ProteoWizard (Chambers et al., 2012) and then processed with TopPIC

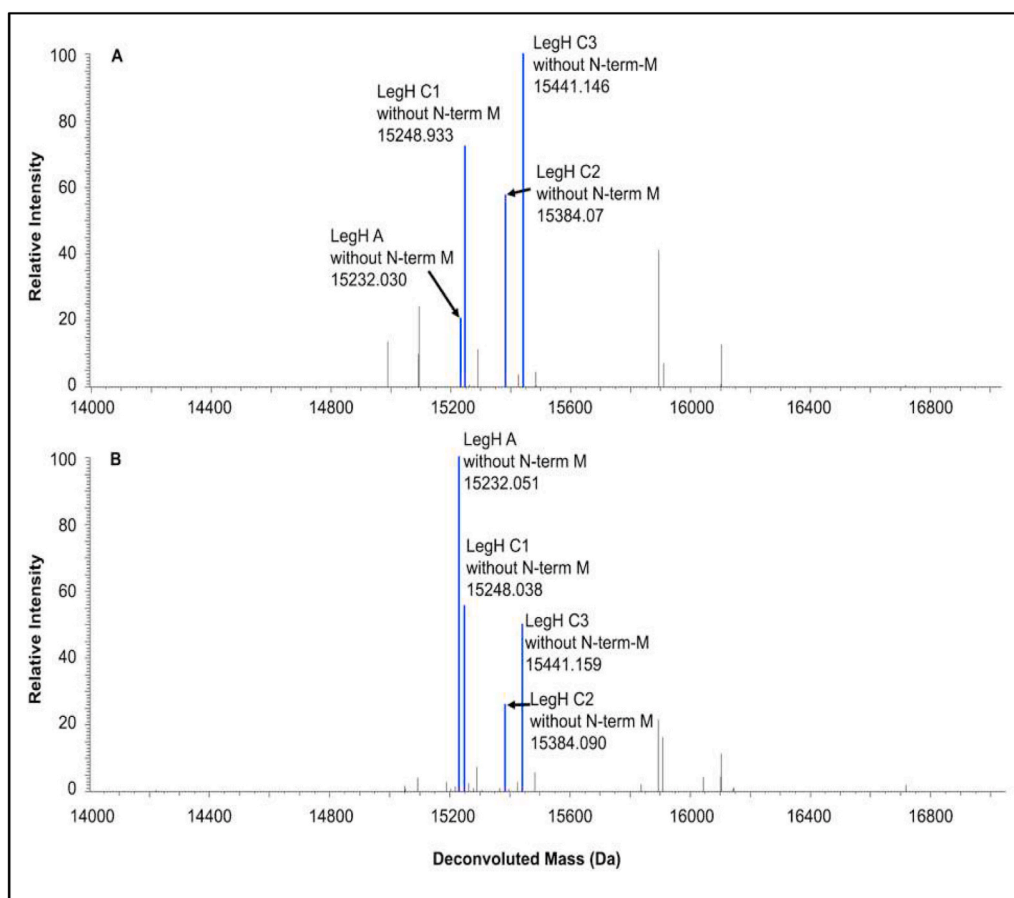


Fig. 1. LegH A, C1, C2, and C3 purified from soy root nodule with DEAE (A) or SEC (B). The sequences of the isoforms were obtained from UniProt (identifiers are P02238, P02235, P02236, and P02237, respectively). The deconvoluted mass of all four isoforms can be matched to the calculated mass from sequences without N-terminal methionine residue.

software package (version 1.3.1) (Kou et al., 2016). The converted spectral files were then loaded into TopFD (Kou et al., 2016) for deconvolution, and the following parameters were utilized: maximum charge state was set to 30, and the S/N ratio cut off for MS1 and MS2 was set to 3.0 and 1.0, respectively. The precursor window size was set to 3.0 m/z, and the m/z error was set to 0.02. The maximum output mass was set to 100,000 Da. The deconvoluted spectral data generated by TopFD were then loaded into TopPIC, with the following parameters for database search: a FASTA database containing the protein sequences of all four soy LegH isoforms (LegH A/C1/C2/C3) were used, and a corresponding decoy database was generated to evaluate the false discovery rate (FDR) of the result. The cut off values of FDR were set to 0.01 both spectrum and proteoform levels. To probe the potential modification on each proteoform, the maximum number of mass shifts was set to two and the search window was set to ± 500 Da. Default values were used for all the other parameters.

2.5. Characterization of heme cofactor

2.5.1. Sample preparation for the heme cofactor analysis

To extract the heme cofactor from LegH, one hundred and 50 μ L of LegH Prep 2.0 was mixed with one hundred and 50 μ L of 20% (v:v) acetic acid in ethyl acetate, vortexed for 5 min, and centrifuged 5 min at 14,000 $\times$ g. One hundred microliters of the top organic layer was carefully transferred into new Eppendorf tubes. The collected organic extraction was diluted 5 $\times$ by adding four hundred microliters of 20% acetic acid in ethyl acetate.

2.5.2. Liquid chromatography with tandem mass spectrometry (LC-MS/MS) data acquisition

Five microliters of the extracted heme cofactor solution was loaded on ACQUITY UPLC CSH C18 column (130 $\text{\AA}$, 1.7 μ m, 1 mm $\times$ 100 mm, Waters), which was maintained at 40 $^{\circ}\text{C}$ and kept at 0.1 mL/min constant flow rate. The elution gradient program was isocratic at 10% B for 2 min, 10%–95% B in 6 min, isocratic at 95% B for 5 min, and isocratic at 10% B for 2 min. Q Exactive Plus Orbitrap (Thermo Fisher) employed a full MS scan followed by four targeted MS² acquisitions. The full MS scan parameters were 140,000 resolution, AGC (Automatic Gain Control) target of 3e6 counts, maximum fill times of 200 ms, and scan range of 250–1400 m/z. Four targeted Parallel Reaction Monitoring (PRM) MS² scans were then performed with the following parameters: 35,000 resolution, AGC target of 2e5 counts, maximum fill times of 100 ms, isolation windows 3.0 m/z, and the normalized collision energy were set to 30, 40, 50, 60 for each PRM scan, respectively. The inclusion list contains only 616.176 m/z, which is corresponding to the chemical formula of oxidized heme B ($[\text{C}_{34}\text{H}_{32}\text{O}_4\text{N}_4\text{Fe(III)}]^+$).

2.5.3. Data analysis for the heme cofactor

The MS1 mass spectra of the heme cofactor were manually extracted within the Qual Browser of the Xcalibur 4.1 software (Thermo Fisher), and the high-resolution MS1 isotopic peaks were compared to the simulated peaks of the chemical formula $[\text{C}_{34}\text{H}_{32}\text{O}_4\text{N}_4\text{Fe(III)}]^+$. Peaks in MS2 mass spectra were manually matched to the possible fragment structure of the heme B.

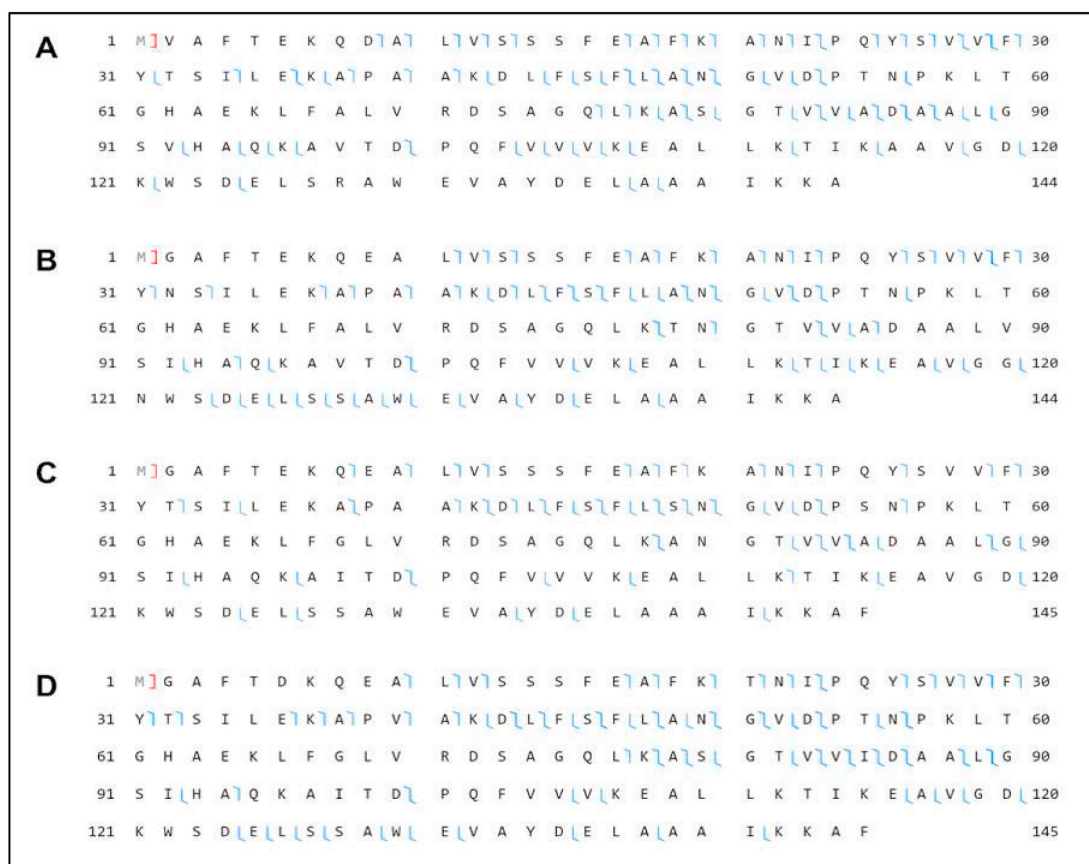


Fig. 2. Mapping of fragment ions from the major proteoform on the sequence of each LegH isoform. The line separators between the residues denote the b- or y-ions identified in the MS2 spectrum. **A.** LegH A isoform; **B.** LegH C1 isoform; **C.** LegH C2 isoform; and **D.** LegH C3 isoform.

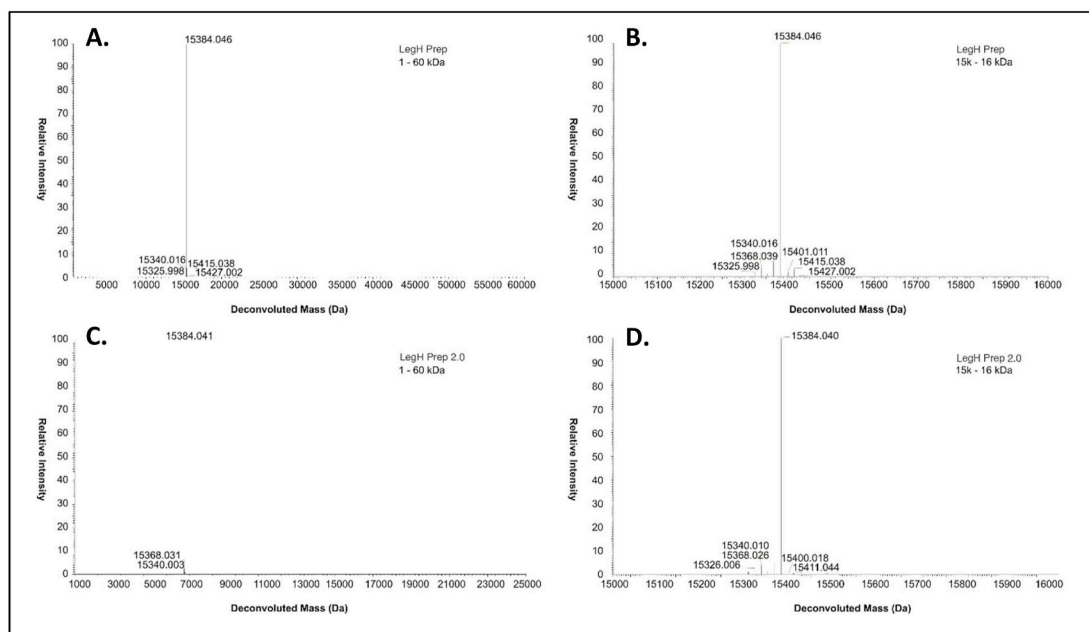


Fig. 3. Deconvolved spectra of LegH Prep and LegH Prep 2.0, showing views of 1 to 60 kDa mass range and zoomed views of 15 and 16 kDa mass range. **A.** and **B.** show the monoisotopic masses for the LegH protein expressed from LegH Prep, **C.** and **D.** show the monoisotopic masses from LegH protein expressed in LegH Prep 2.0.



Fig. 4. Mapping of identified peptides of LegH over the protein sequence.

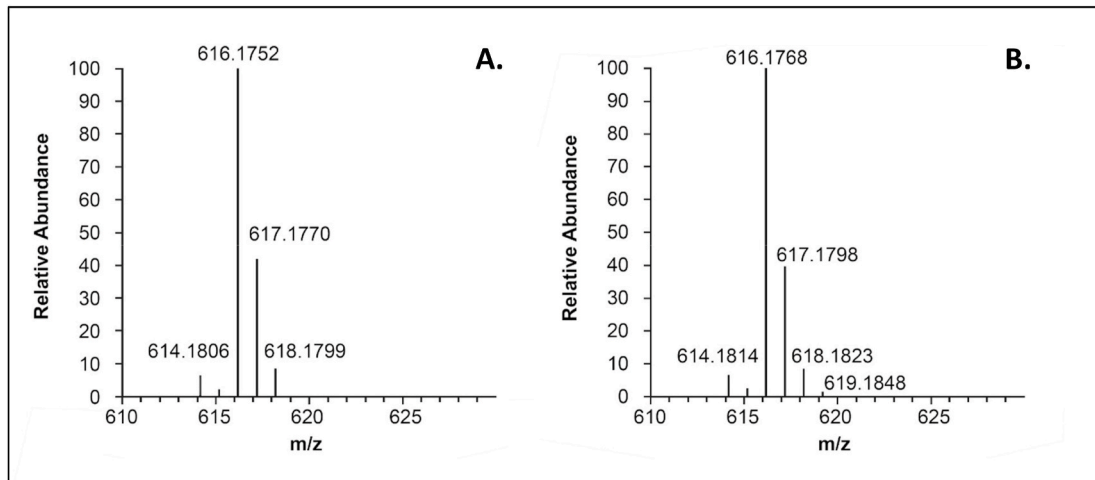


Fig. 5. The MS1 spectra of Heme. A. The high-resolution mass spectrum of the heme cofactor extracted from the LegH Prep. B. The simulated heme B isotopic peaks.

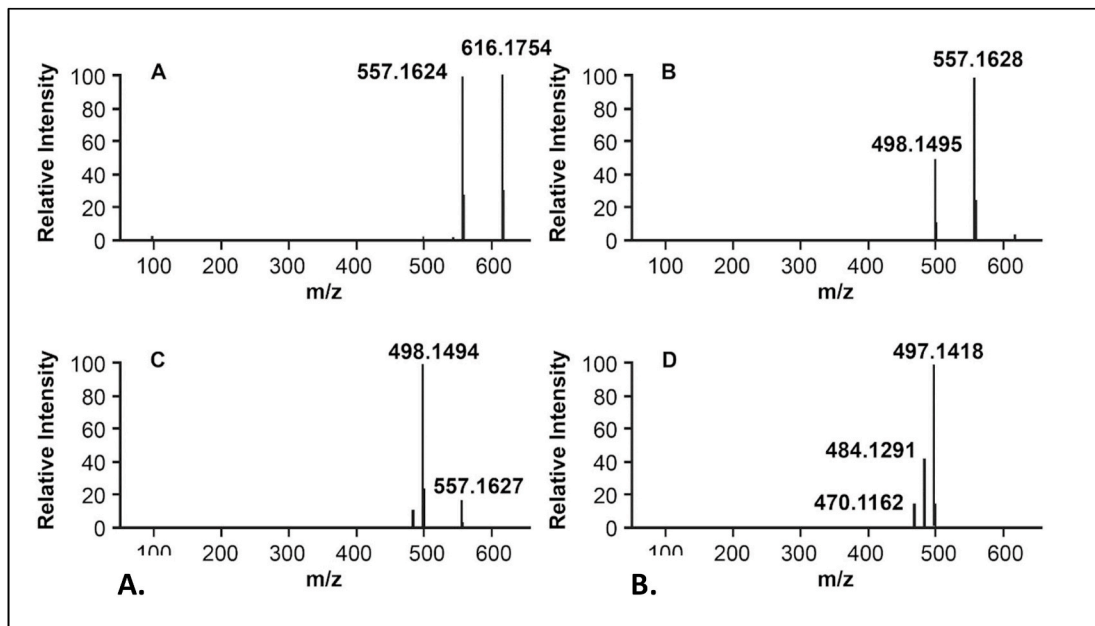


Fig. 6. The MS2 spectra of the heme cofactor precursor under various normalized collision energies. A. NCE = 30; B. NCE = 40; C. NCE = 50; D. NCE = 60.

2.6. Bioinformatics search

2.6.1. *Pichia* protein sequences

Potential allergenicity and toxicity of the *P. pastoris* proteins present in LegH Prep 2.0 at $\geq 1\%$ of the total protein fraction were evaluated consistent with Codex recommendations (Alimentarius, 2009; FAO/WHO, 1996). A protein would be included to further

bioinformatics search if it were present at an abundance higher than 1% in any batch based on the shot-gun proteomics data. The protein sequences were based on the genome sequence of *P. pastoris* CBS7435 (Küberl et al., 2011).

2.6.2. AllergenOnline (AOL) search

The AllergenOnline (AOL) database (<http://www.allergenonline>).

Table 1

List of possible structures of the product ions from heme cofactor fragmentation.

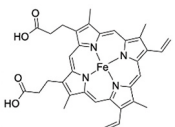
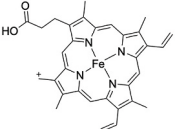
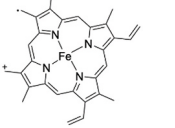
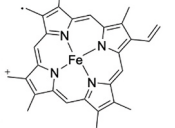
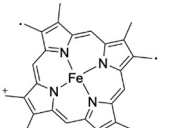
Mass (Da)	Normalized Collision Energy	Chemical Structure	Comment
616.1754	30		Heme B precursor ion
557.1628	40		Loss of -CH ₂ COOH side chain (59.01 Da) from heme B
498.1495	40		Loss of another -CH ₂ COOH side chain (59.01 Da) from 2
484.1291	50		Loss of CH ₂ = side chain (14.02 Da) from 3
470.1162	60		Loss of CH ₂ = side chain (14.01 Da) from 4

Table 2List of seven *Pichia* proteins from the LegH Prep 2.0.The proteins shown here are proteins from *Pichia* sp. that represent $\geq 1\%$ of the total protein fraction in at least one batch of LegH Prep 2.0.

	Accession	Protein
1	XP_002493171.1	NAD(+)-dependent formate dehydrogenase
2	XP_002490864.1	Translational elongation factor EF-1 alpha
3	XP_002493194.1	Uroporphyrinogen decarboxylase
4	XP_002490697.1	Hexokinase-2
5	XP_002491701.1	Acetate-CoA ligase
6	XP_002491345.1	Glyceraldehyde-3-phosphate dehydrogenase, isozyme 3
7	XP_002493065.1	Transketolase

org/) is a peer-reviewed allergen list and sequence database that is updated each year (Goodman et al., 2016). AOL (2020 release, v20) was used as the primary tool in searching for matches to allergenic sequences. Three comparison methods were used: an overall FASTA3 full-length search (using E-scores of 1.0), a sliding 80 amino acid (AA) window FASTA search (E-score of 10, identity > 35%), and an exact word search tool looking for 8 AA exact matches. The E-score criteria limits have been established on the AllergenOnline database based on extensive testing to limit irrelevant matches while identifying alignments that may cause cross-reactivity (Goodman et al., 2016).

2.6.3. BLASTP in NCBI entrez protein database

The BLASTP search algorithm is available on the NCBI website (<http://www.ncbi.nlm.nih.gov/BLAST>). Protein sequence entries in the Entrez search and retrieval system are compiled and maintained by the NCBI of the National Institutes of Health (U.S.A.). The database is updated or modified every few days. Sequences are annotated with

publications or with notation of likely protein types; much of that is added by computer algorithm and some simply suggest minor similarities to a previously identified allergen type. Since AllergenOnline (AOL) is updated once per year, the NCBI protein database was searched using keyword limits ("allergen", "toxin" and "toxic") to identify possible matches to proteins that may represent newly discovered allergens or toxins. The searches for matches between newly deposited proteins and the LegH and the seven host proteins were performed in August 2020. Searches were also performed for each sequence without any keyword to evaluate whether similar proteins exist in other organisms that might provide information of safe exposure to homologous proteins. The additional searches ensured that sequences, which might not yet have been entered into AOL v20, were included in the analysis (Siruguri et al., 2015). Additional bioinformatics searches with NCBI and BLASTP were used to compare sequence identity ranking with proteins of risk or from known safe food sources. E-score values, the length of the alignments, and the identity percentage of any identified match from the BLASTP results were evaluated to judge the significance of any alignment.

2.7. In-vitro pepsin digestibility study of LegH

The test system for this study was an *in vitro* digestion model using pepsin in simulated gastric fluid (SGF). The SGF preparation and digestion procedures were based on the methods described by Thomas et al., 2004 (Thomas et al., 2004) as modified by Ofori-Anti et al., 2008 (Ofori-Anti et al., 2008). In brief, a 1.6 mL digestion reaction contained 400 µg target protein (the LegH Prep), 4000 units of pepsin, and 3.15 mg NaCl, at pH = 2.0. Reactions were incubated at 37 °C until the designated time points by adding 35 µL quenching solution (200 mM NaHCO₃, pH = 11) and heating at 95 °C for 10 min. The digestion products were analyzed on SDS-PAGE.

3. Results

3.1. Analysis of the LegH purified from soy root nodule

To confirm the identity of the soy LegH Prep 2.0 protein, we first conducted a mass spectrometry-based characterization of the native LegHs from the soy root nodule. To determine whether LegH C2 protein expressed in *Pichia* was equivalent to native LegH C2, we first purified LegH isoforms from soy root nodules to examine their chemical composition. Two samples were obtained, one collected from DEAE and one collected from SEC. Both were analyzed using the protocol described in 2.3. In both batches, all four isoforms of the LegH were identified by exact mass matches in both batches (Fig. 1).

3.1.1. Top-down proteomics analysis of LegHs from soy root nodule

Next, to characterize all four isoforms of soy LegHs, we performed a top-down proteomics study. All four isoforms were identified in each purified sample from the soy root nodule, and for each of them, the major proteoform identified was the unmodified versions, which was consistent with the result from the IPMS experiment. The matched MS2 fragment ions for the major proteoform were mapped over the sequence of each isoform (Fig. 2). Detailed information can be found in Supplemental Table S3-S7.

3.2. Intact protein mass spectrometry (IPMS) analysis of LegH expressed in *pichia*

We conducted IPMS to assess whether the LegH protein manufactured with the new process (LegH Prep 2.0) was equivalent to LegH protein manufactured from the original process (LegH Prep). After deconvolution, for both samples the mono-isotopic mass of the only major peak was reported as 15384.04 Da (Fig. 3), which can be matched to soy LegH C2 with *N-terminal* methionine excision (calculated mass = 15384.0399 Da) and also showed consistency with the LegH C2 isolated

Table 3Amino acid sequences of the seven *Pichia* proteins.

Organism	Protein	Native protein sequence
Komagataella phaffii (P. pastoris)	NAD(+)-dependent formate dehydrogenase, GI:254572123, ACC: XP_002493171.1	MKIVLVLYSAGKHAADPEPKLYGCINELGIRQWLEKGGHELVTTSDEKEGENSELEKHIPDADVIISTPFPHPAYITK ERIQAKAKKLKLLVAVGVGSDHIDLIDYEQNGLDISVLEVTGNSVVSVAEHVVMITLNLVRNFPVAHEQIVNHGWD VAAIAKDAYDIEGKTATITAGRIGRYRVLRLVAFNPKELLYDYQGLPKAEAEKVGARRVDTVEELVAQADVTV NAPLHAGTKGLVKNELLSKFKKGAWLNTARGAICNAQDVADAVASGQLRGYGGDVWFQPPAPKDPHPW RDMRNKYGYGNAMTPHYSGTTLDAQVRYAEGTKNLNSFLTCKFDYRQDVILLNGKYTKAYGNDKKVA
Komagataella phaffii (P. pastoris)	Translational Elongation factor EF-1 alpha, GI:254567507, ACC: XP_002490864.1	MGKEKLHVN VV VIGHVDAGKSTTTGHLIYKCGGIDKRTIEKFEKAEELGKGSFKYAVWLDKLAERERGITIDIALWKFETPKY HVTVIDAPGHRDFIKNMTGTSGADCAILVIASGIGEFAGISKDGQTRHALLFTLGKQLIVAINKMDSVKWSQ KRYEEIVKETSNIKKVGNPNKTPVPFVPISGWNGDNMIEPSSNCDWYKWEKETKAGGATKGTLLAIDSIDPPSR PTDKPLRLPLQDVYKIGGIGTVPVGRVETGVKAGMVTTFAPAGVTTEVKSVMEMHHEQLEQGVPGDNVGFNVKN VSVKEIRRGNVCGDSKNDPPKAAESFNAQVILNHPGQISAGYAPVLDCHTAHACKFDELIEKIDRRTGKKTENPK FIKSGDAAIVKLVPSPKMCVEAFTDYPLGRFAVRDMRQTVAVGVKSVVTKDAGKVTKAAQKAACK
Komagataella phaffii (P. pastoris)	Uroporphyrinogen decarboxylase, GI:254572169, ACC: XP_002493194.1	MSRFPPELKNLILRAARGEKVERPPIWIMRQAGRYLPEYHEVKGGRDFETCRDAEIASIETIPITHFGLIDAAIIF SDILVIPQAMGMEVKMVDKVGQFPNPLRKPSDLHLKLDVVLKELDWAFKAISLTRKKNLNGRVPPLGFCGAPWTL LVYMTDEGGGTNLMFRFAKEWYKFTKESHQQLQITDVAVEFLAQVVGAGQMLQVFESWGELGPDEFDEFSLPYLR QISSKLPLRLKELGITENVPIVFAKGSWYALEQLCDSGYDVVSLDWLFRPSDAVQIANGRIALQGNLDPGTMYGSK ETISKKVDDKMIKFGGQKQNYIIFGHGTHPFMDPEQIRWFLQECHRIGSQ
Komagataella phaffii (P. pastoris)	Hexokinase-2, GI: 254567173, ACC: XP_002490697.1	MPIAKPANQVASTITTSQHLSDRVLIEQDAFEVSPTKLQQIVAHFVEELKGLSAGKGNIPMIPVWVMDYPNGTETG DYLAIDLGGTNLRVVLVHLLGGQKFTEQKYLHLPKGMRTTRNRDELFEFIADCKLEFFYKLHPNGIEKGTLLPLG FTFSPASQTRIDTGLVLRWTKGFDIPNVEGEDVVPPLMDKINEKLPKIKVVALINDTAGALVASRYTDPTEMGILF GTGVNGAYYDRVNGIEKLKGLLDPITDESPMLINCEYGSFNDNEHESLPRTKYDILIDEQSPRPGQQAPEKMTAGYYL GELIRLILVELYEEKQVQKYSKDESEKILLYTPYLLDTSFLAEIEGDQDLNENPEVVRFLQEFELKIEPTLDERRLSRLS EIIGNRSARLSVCGIGAACTKMNKKCHCAADGVSFHKYKPFPERAADSLADIFGWKEQNIQPKNYPIQIVPSQDGS VVGAIAALAHARQAKGLSLGLAEAYKKA
Komagataella phaffii (P. pastoris)	Acetate-CoA ligase, GI: 254569182, ACC: XP_002491701.1	MPLDNEHLLHENSIDPPKGFERHPGTPNIPGGWEEYLKLYNQSINPSKFFGEKAKEFLSWATPFTDARYPPGNGFQ NGDSAAWFLNGELNASYNCDRHALKNPDKPAIYEADPNQGRVTYVYELLGDCRIAQVLTDLGVKKGDTVA VY LPMVPEAITTLAIVRIGAIHSVVFAGFSAGSLDRDLADSRIVTTSDESRLGGKIETKKIVDEALKSCPDVRNVLVFKR TGTPHLPWVEGRDLWWHEEIIHKVPYSPVNVRSDETSFLYTSGSTGKPKIQHSTAGYLLGALLTTKYVFDVQ GDDILFTAGDVGWITGHSYVYVYGPLLNGATTVVFEFTPAFYPDYSRYWDIVDKHKVTQFYVAPTALRLKRAKS KYVQNHLDSIRIVLGSGVEPIAAEVWEWYNEYVGRGKAHICDTYWTETGSHIIAPIAGVSKTKPGSASFPPFGIDP VILDATTGEELKGNVNEGVLAIKPNWPSMARTVWKDYNRFLDTYLRPEYGYFTDGAARDQEGFYVVLGRVDD VVNVSGHRLSTAIEIESALIEHNLVGESAVVGFPDELTSASAAVFLSKDDVNDNPAEVKKELILTVRKEIGFPAAPKLI VSDLPKTRSGKIMRRLRKVLAGEEDSLGDIETLSNPSIVEEIIISTVKRDARK
Komagataella phaffii (P. pastoris)	Glyceraldehyde-3-phosphate dehydrogenase, isozyme 3, GI: 254568470, ACC: XP_002491345.1	MAITVINGFGFRGRLVLRVALSRADIKVVAINDPFAIPEYAAVMFYKDYSTHAKYKGEVSASGNKINIDGNEITVFQE RDPVNPVWKGAGVDYVIESTGVFTTLEGAQKHIDAGAKKVVITAPSKAMPFVVGVEEYKTSDLNIVNSASCTTNC LAPLAKVNDITFGIESGLMTTVHSMATATQKTVDGSPSHKDWRRGTASGNIIPSTGAAGAVGKVPGLNGLKLTGLA FRVPTVDVSVVDLTNLNKETTYEEIKSVIKAASEGKLKGVLYTEDAVVSSDFLGDERSIFDASAGIQLTPSFVKLI SWYDNEYGYSTRVVDLLQHVAKA
Komagataella phaffii (P. pastoris)	Transketolase, GI: 254571911, ACC: XP_002493065.1	MARIPKAVSYNDIIDLVIKTRFCYVLDLVEQYGGGHPGSAMGMVAIGIALWKYQMKYAPNDPDYFNDRFVLSN GHVCLFQYLFQHLTGLKEMTVKQLQSYHSSDYHSLTPGHPEIENPAVEVTTGPGQGGSNAVGMAIGSKNLAATY NRPFGFVVDNTIYAIVGDAQLQEGPALESISLAGHLALDNLIVYDNNQVCCDGSVDVNNNTEDISAKFRAQNNVNI EVENGSRDVA TLVKAIEWAKAENERTLINVRTEIGQDSAFGNHHAHAGSALGEEGIRELKAKYGFVDARKFWFP QEVYDFFAEKPAEGDQLVANWKKLLDEYVKNYPQEGEELKARIRGELPKNWFSEFIPQDKPTEATRTSAREIVRS LGQNLQVIAGSGDLSVILLSNWGGVGYFFNPKLQTFCLGGDYSGRYIEFGIREHSMCAIANGLAAYNKGFTLPIT STFYMFYLYAAPALRMAALQELKAIHIATHDSIGAGEDGPTHQPIALSSLRAMPNFYIRPADATEVAALFEVAVE LEHSTLFSLSRHEVEQYPGKTSAGAKRGYVVEDCEGKPDVQLIGAGSELEFAVKTARLLRQKQGWKVRVLSFPC QRLFDQQLSLAYRRSVLRREGEVPTTVVVEAYVAYGWERYATAGYTMNTFGKSLPVEDVYKYFYGYTEPKIGEKVAA VNSIKASQILYEFTDLKGKPKHDKL

from root soy nodule (Fig. 1). There is no other high-intensity peak observed within the mass range of 60 kD, though some minor peaks were observed in LegH Prep 2.0, which is consistent with the mass spectrometry analysis from the original LegH Prep (Fraser et al., 2018; Jin et al., 2018). Although the identities of these minor peaks were not fully characterized, based on their mass shifts, they are likely peptide fragments hydrolyzed from the LegH. Neither unexpected glycosylation nor other significant modifications was observed on the LegH expressed in LegH Prep 2.0.

3.3. Shotgun proteomics of the LegH samples

To investigate the residual *Pichia* host proteins, mass spectrometry-based quantitative shotgun proteomics was utilized to identify and quantify the proteins in the LegH Prep 2.0 using three independent production lots. To estimate the relative abundance of identified

protein, the total LFQ intensity of each protein from a single sample was normalized to the total intensity of the identified proteins in the sample. To evaluate the allergenicity and toxicity of the host cell proteins, all the proteins from *P. pastoris* with the relative abundance greater than 1% were subjected to a detailed bioinformatics analysis, according to the method described in 2.2. Results are discussed in Section 3.5.

3.3.1. Peptide mapping of the LegH from the soy LegH from pichia

To evaluate the consistency between the LegH yielded from LegH Prep 2.0 and the desired sequence, peptide mapping was performed on the identified peptides in the shotgun proteomics analysis (Fig. 4). In total, there were 21 peptides from LegH identified (Supplemental Table 2), the overall sequence coverage was 97.9% with most of the residues covered, except the N-terminal methionine and two C-terminal residues following Lys143. The result showed good consistency with the designated sequence for the recombinant LegH expression, no mutation or

Table 4Summary of sequence alignments of the seven *Pichia* proteins in Table 3.

Protein Name	GeneInfo Identifier	Accession	No. of AA ^a	AOL ^b Matched Allergen ^c	AOL Best ID ^d	<i>Saccharomyces</i> sp. Best ID
NAD(+) dependent formate dehydrogenase	254572123	XP_002493171.1	365	n/a	n/a	61%
Translational elongation factor EF-1 alpha	254567507	XP_002490864.1	459	n/a	n/a	89%
Uroporphyrinogen decarboxylase	254572169	XP_002493194.1	361	n/a	n/a	74%
Hexokinase-2	254567173	XP_002490697.1	496	n/a	n/a	58%
Acetate-CoA ligase	254569182	XP_002491701.1	668	n/a	n/a	67%
Glyceraldehyde-3-phosphate dehydrogenase, isozyme 3	254568470	XP_002491345.1	333	Tri a 34.0101	70%	81%
Transketolase	254571911	XP_002493065.1	707	Pen o 18.0101	21%	45%

n/a: not available or no answer.

^a AA: amino acids.^b AOL: AllergenOnline.^c Allergen name in IUIS allergen list unless denoted as unassigned by IUIS.^d ID: identity (%).

deletion was observed.

The red color-coding denotes the cleavage sites of the trypsin protease, and the black bands show the position of the identified peptides.

3.4. Characterization of heme cofactor

To discern the chemical identity of the heme cofactor in the LegH protein from *Pichia*, tandem mass spectral analyses was performed and compared to the chemical structure of heme B. Results of these analyses are shown below. The MS1 spectra of heme showing the observed isotope distribution matching the expected one. MS1 spectra collected within the retention time window between 8.5 and 9.5 min were extracted manually (Fig. 5). Panel A is the experimental spectrum within the mass window of 610 Da and 630 Da, and Panel B is the calculated spectrum from the chemical formula of heme B (C₃₄H₃₂O₄N₄Fe). The measured mono-isotopic mass of heme B was 616.1752 Da, the error between the experimental value and theoretical mass is +2.60 ppm. All the isotopic peaks of heme B can also be matched accurately with the calculated masses.

3.4.1. The fragment ions in the MS2 spectra of heme cofactor

The MS2 spectra were collected for the heme cofactor precursor with the targeted mass of 616 Da. Four different normalized collision energies (NCE) were applied in high-energy dissociation (HCD) for precursor fragmentation: 30, 40, 50, and 60. The masses of product ions were manually matched to the fragments of the prosthetic group heme B (Fig. 6, Table 1). As shown in Table 1, the identified fragment ions matched the theoretical fragments produced from heme B. Results of these analysis demonstrated that the biosynthesized heme cofactor from LegH Prep. 2.0 matches that of heme b and excludes other porphyrin species.

3.5. Bioinformatics analysis of sequence homology

We assessed the potential allergenicity and toxicity of the *Pichia* sp. proteins present in LegH Prep 2.0 at ≥1% of the total protein fraction, consistent with Codex recommendations. Results from the quantitative shotgun proteomics data on three independent production lots of the LegH Prep 2.0 indicates that in addition to LegH, seven new proteins were present at 1% or above in at least one batch (Table 2). A table showing the relative abundance of the 7 proteins in each batch is provided in Supplemental Table S8. It is important to note that a total of 11

proteins were originally identified; however, 4 were previously evaluated and determined to be unlikely to pose risk of allergenicity or toxicity (Jin et al., 2018). The remaining seven novel proteins are: NAD (+)-dependent formate dehydrogenase (XP_002493171.1), Translational elongation factor EF-1 alpha (XP_002490864.1), Uroporphyrinogen decarboxylase (XP_002493194.1), Hexokinase-2 (XP_002490697.1), Acetate-CoA ligase (XP_002491701.1), Glyceraldehyde-3-phosphate dehydrogenase isozyme 3 (XP_002491345.1), and Transketolase (XP_002493065.1). The amino acid sequences of the seven proteins are listed in Table 3. The bioinformatics analysis described in this study used the same methodology that was used in GRAS Notice 737 (U.S. Food and Drug Administration, 2018) to evaluate potential allergenicity and toxicity; results are summarized in Table 4.

AOL search tool results of the seven identified proteins are shown in Tables 5–11. In summary, bioinformatics data of the proteins (Tables 5–11) showed that no significant matches were found; with the exception of one, which matched to the highly evolutionarily conserved GAPDH from wheat (*T. aestivum*) protein. Sequence homology data shows that there are very good alignments to homologous proteins from many molds including *Saccharomyces* sp., *A. oryzae*, and others (Tables 5–11). Thus, the risks from consumption of these proteins appear to be minimal. In the case of GAPDH, there were alignments to proteins identified as being from toxic organisms, but only as common homologues. GAPDH is a ubiquitous protein whose homolog can be also found in *A. oryzae*, *P. roqueforti*, and human (Table 10). Identity scores were higher for commonly consumed, non-allergen foods like *S. cerevisiae* compared to *T. aestivum* (Fig. 7). Therefore, the likely risk of allergenicity and toxicity associated with *Pichia* GAPDH is relatively low.

3.6. In vitro pepsin digestibility

LegH Prep 2.0 was subject to *in vitro* pepsin digestion in simulated gastric fluid (SGF) at pH 2.0. Samples were taken over a 60 min time course. All proteins, including LegH and *Pichia* proteins were digested to non-detectable levels in 2 min (Fig. 8). This result suggests no protease resistant proteins were present in LegH Prep 2.0.

4. Discussion

Soy LegH is used to impart the flavor and color of animal meat to

Table 5**Bioinformatics results of NAD (+)-dependent formate dehydrogenase (XP_002493171.1).**

The 365 AA sequence from NAD (+)-dependent formate dehydrogenase (XP_002493171.1) was compared with [AllergenOnline.org](https://www.allergen.org/) version 20 and with NCBI Protein on August 2020.

AOL full FASTA3	- Relatively low-scoring alignments were identified to Putative <i>Pis</i> v 2.0201 allergen 11, with E-scores of 0.05 and identity of 30% on a 100 AA overlap. - Low-scoring alignments were identified to Allergen <i>Sal</i> k 4.02 (<i>S. kali</i>), with E-scores of 0.75 and identity of 21.5% on a 65 AA overlap. - Low-scoring alignments were identified to allergen profiling (<i>S. melongena</i>), with E-scores of 0.72 and identity of 27.4% on a 106 AA overlap. - All three alignments are clearly poor and do not represent conservation of homology. Thus, it is unlikely to represent a risk of cross-reactivity.
AOL 80mer FASTA3	No alignments with >35% identity over any 80 AA match were identified.
AOL Exact 8 AA	No exact 8 AA match was identified.
NCBI BLASTP Allergen	No significant match was identified using BLASTP and keyword “Allergen”.
NCBI BLASTP “Toxin or Toxic” keyword search	The matches were identified to homologous proteins from bacteria (e.g. <i>Bordetella</i> sp., <i>S. aureus</i>, <i>Bacillus</i> sp.), with high identities and significant E-scores. However, the linkage appeared to be due to the ability of the homologous proteins in a wide array of organisms to detoxify exogenous formate. No published evidence was found that these proteins are toxins.
NCBI BLASTP No keyword	Many matches of very high scoring identities with homologous proteins from other yeasts and molds were identified with identities from 42% to 86%.
Conclusion: No significant matches were found in any other searches. It is unlikely that there is any risk of allergy for this protein. The risk of toxicity is similarly low. In addition, there are very good alignments to homologous proteins from many molds including <i>Saccharomyces</i> sp. Therefore, risks from consumption of this protein appear to be minimal.	

Table 6**Bioinformatics summary results of protein XP_002490864.1.**

The 459 AA Translational elongation factor EF-1 alpha (XP_002490864.1) was compared with [AllergenOnline.org](https://www.allergen.org/) version 20 and with NCBI Protein on August 2020.

AOL full FASTA3	Very low-scoring alignments with an E-score of 0.25 were identified to <i>bet</i> v 1. The alignment is unlikely to represent a risk of possible cross-reactivity.
AOL 80mer FASTA3	No alignments with >35% identity over any 80 AA match were identified.
AOL Exact 8 AA	No exact 8 AA match was identified.
NCBI BLASTP Allergen	Many high scoring alignments were identified using allergen and allergy keywords to Translational elongation factor EF-1 alpha from other organisms across different kingdoms. No published evidence was found that these proteins are allergens.
NCBI BLASTP “Toxin or Toxic” keyword search	Many high scoring alignments were identified using allergen and allergy keywords to Translational elongation factor EF-1 alpha from other organisms across different kingdoms. No published evidence was found that these proteins are toxins.
NCBI BLASTP No keyword	Many high scoring identity matches with homologous proteins from other yeasts and molds were identified. The highest scoring alignment with a <i>Saccharomyces</i> sp. was with an E-score of 0.0 and identity of 89% over 459 AA.
Conclusion: No significant matches were found in any other searches. It is unlikely that there is any risk of allergy for this protein. The same is true for toxicity. In addition, there are very good alignments to homologous proteins from many molds including <i>Saccharomyces</i> sp., with 89% identity over 459 AA. Thus, it is very unlikely that this is an allergen or a toxin.	

Table 7**Bioinformatics summary results of protein XP_002493194.1.**

The 361 AA, Uroporphyrinogen decarboxylase (XP_002493194.1) was compared with [AllergenOnline.org](https://www.allergen.org/) version 20 and with NCBI Protein.

AOL full FASTA3	- Low-scoring alignments were identified to defensin 2 (<i>A. hypogaea</i>), with E-scores of 0.13 and identity of 30.4% on a 46 AA overlap. - Low-scoring alignments were identified to Putative allergen protein (<i>L. bostrychophila</i>), with E-scores of 0.91 and identity of 31.1% on a 61 AA overlap. - Low-scoring alignments were identified to <i>Vigna</i> <i>Vig</i> r 6 Cytokinin-specific binding protein (<i>V. radiata</i>), with E-scores of 0.35 and identity of 36.7% on a 49 AA overlap. - The alignment is clearly poor and does not represent conservation of homology. Thus, it is unlikely to represent a risk of cross-reactivity.
AOL 80mer FASTA3	No alignments with >35% identity over any 80 AA match were identified.
AOL Exact 8 AA	No exact 8 AA match was identified.
NCBI BLASTP Allergen	No significant match was identified using BLASTP and keyword “Allergen”.
NCBI BLASTP “Toxin or Toxic” keyword search	Many matches were identified to homologous proteins from bacteria that are associated with toxicity (e.g. <i>Bordetella</i> sp., <i>S. aureus</i>, <i>Bacillus</i> sp.). No published evidence was found that these proteins are toxins.
NCBI BLASTP No keyword	Many high scoring identity matches with homologous proteins from other yeasts and molds were identified, including those from <i>S. cerevisiae</i> (E-score 0.0, identity 72% over 361 AA).
Conclusion: No significant matches were found in any other searches, and based on the results, it is very unlikely that there is any risk of allergen or toxicity for this protein.	

Table 8**Bioinformatics summary results of protein XP_002490697.1**

The 496 AA Hexokinase-2 (XP_002490697.1) was compared with [AllergenOnline.org](https://www.allergen.org/) version 20 and with NCBI Protein.

AOL full FASTA3	- Low-scoring alignments were identified to <i>Pru</i> p 7 Peamaclein (<i>V. radiata</i>), with E-scores of 0.35 and identity of 36.7% on a 49 AA overlap. - The alignment is clearly poor and does not represent conservation of homology. Thus, it is unlikely to represent a risk of cross-reactivity.
AOL 80mer FASTA3	No alignments with >35% identity over any 80 AA match were identified.
AOL Exact 8 AA	No exact 8 AA match was identified.
NCBI BLASTP Allergen	Low score alignment was identified to <i>Per</i> a 2 allergen aspartic protease precursor from American cockroach, <i>P. Americana</i> with E-score of 0.28. The alignment is unlikely to represent a risk of possible cross-reactivity.
NCBI BLASTP “Toxin or Toxic” keyword search	No significant match was identified using BLASTP and keyword Toxin or Toxic.
NCBI BLASTP No keyword	Many high scoring identity matches with homologous proteins from other yeasts and molds were identified, including those from <i>S. cerevisiae</i> (E-score 0.0, identity 58% over 476 AA) and <i>A. oryzae</i> (E-score 2e-162, identity 52% over 474 AA).
Conclusion: No significant matches were found in any other searches, and it is very unlikely that there is any risk of allergen or toxicity for this protein. There are very good alignments to homologous proteins from many molds including <i>Saccharomyces</i> sp. Therefore, risks from consumption of this protein appear to be minimal.	

plant-based meat alternatives. A LegH Prep from *P. pastoris* was previously assessed for potential allergenicity and toxicity (Fraser et al., 2018; Jin et al., 2018). These data were published in a FDA GRN 737 (U.S. Food and Drug Administration, 2018). Recently, a processing change was introduced to the production of LegH Prep (LegH Prep 2.0), to lower

Table 9**Bioinformatics summary results of protein XP_002491701.1.**

The 668 AA Acetate-CoA ligase (XP_002491701.1) was compared with [AllergenOnline.org](https://www.allergenonline.org) version 20 and with NCBI Protein.

AOL full FASTA3	• No alignments with E < 1.0 were identified.
AOL 80mer FASTA3	• No matches were identified with >35% identity in any 80 AA match.
AOL Exact 8 AA	• Two exact 8 AA matches were identified to <i>Cha o</i> 3, the glycosyl hydrolase 5 family protein from <i>C. obtusa</i> (Japanese cypress).
NCBI BLASTP Allergen	• Two modest scoring alignments were identified using allergen or allergy keyword to fengycin synthetase <i>FenA</i> from <i>B. subtilis</i> (E-score 1E-21, identity 25% over 557 AA) and hypothetical protein FF38_10715 from <i>L. cuprina</i> (E-score 2E-16, identity 23% over 551 AA). No published evidence was found that these proteins are allergens.
NCBI BLASTP "Toxin or Toxic" keyword search	• Many high scoring alignments were identified using toxin and toxic keyword to acetate-CoA ligase from other organisms across different kingdoms. The toxin or toxic keyword is associated to the toxic bacteria or toxic planktons. No published evidence was found that these proteins are toxins.
NCBI BLASTP No keyword	• Many high scoring identity matches with homologous proteins from other yeasts and molds were identified, including those from <i>A. oryzae</i> (E-score 0.0, identity 65% over 665 AA) and from <i>P. roqueforti</i> (E-score 0.0, identity 65% over 663 AA).
Conclusion: Acetate-CoA ligase (XP_002491701.1) is unlikely to pose any risk of allergenicity or toxicity. In addition, there are very good alignments to homologous proteins from many molds including <i>A. oryzae</i> and <i>P. roqueforti</i> . Thus, it is very unlikely that this is an allergen or a toxin.	

production costs and further improve performance. Impossible Foods' engineered *Pichia* production strain complies with the Organization for Economic Co-operation and Development (OECD) criteria for Good Industrial Large Scale Practice (GILSP) microorganisms (OECD, 1992). It also meets the criteria for a safe production microorganism as described

Table 10**Bioinformatics summary results of protein XP_002491345.1.**

The 333 AA GAPDH, isozyme 3 (XP_002491345.1) was compared with [AllergenOnline.org](https://www.allergenonline.org) version 20 and with NCBI Protein.

AOL full FASTA3	• One highly significant match with 3.2E-97 E-value and 70% identity over 332 AA was identified to a putative allergen <i>Tri a</i> 34.0101, a GAPDH from <i>T. aestivum</i> , identified as an allergen based binding IgE in baker's asthma (Sander et al., 2011).
AOL 80mer FASTA3	• There was no published allergic reaction to ingestion of <i>Tri a</i> 34.0101. <i>P. camemberti</i> is widely consumed with brie and other soft-ripened cheeses (Galli et al., 2016) and <i>A. oryzae</i> is widely used in Asia as koji in soy sauce and soy paste fermentation (Kitamoto, 2015), both of which have a GAPDH that is closer to the allergenic ones than is the <i>Pichia</i> protein XP_002491345.1 (Fig. 7).
AOL Exact 8 AA	• XP_002491345.1 has a very high sequence similarity to the GAPDH in <i>S. cerevisiae</i> with alignment E-score of 0.0 and 81% sequence identity. Therefore, cross-reactivity between XP_002491345.1 and <i>Tri a</i> 34.0101 is highly unlikely and the risk of allergy is likely to be very low.
NCBI BLASTP Allergen	• One alignment was found with >35% identity over 80 AA. It was the same a GAPDH from <i>T. aestivum</i> as the full-length FASTA, with highest scoring 80mers of 87.5% identity.
NCBI BLASTP "Toxin or Toxic" keyword search	• A number of 8 AA matches were identified to the same glyceraldehyde-3-phosphate dehydrogenase from <i>T. aestivum</i> as the full-length FASTA.
NCBI BLASTP No keyword	• Four high identity matches were found to GAPDHs of <i>P. Americana</i> , with an E-score ranging from 7e-160 to 0.0 and identities from 66% to 83%.
Conclusion: There is a minimal risk of allergy based on the alignment with GAPDH from <i>T. aestivum</i> . Thus, the risk of allergy is likely to be very low. There were alignments to proteins identified as being from toxic organisms, but only as common homologues. This is a widely spread conserved protein whose homolog can be also found in <i>A. oryzae</i> (E-score 6E-165, identity 67% over 329 AA), from <i>P. roqueforti</i> , and from human with E-score of 2E-165 and 68% sequence identity. Thus, the likely risk of toxicity is low.	

Table 11**Bioinformatics summary results of protein XP_002493065.1.**

The 707 AA Transketolase (XP_002493065.1) was compared with [AllergenOnline.org](https://www.allergenonline.org) version 20 and with NCBI Protein.

AOL full FASTA3	• No alignments with E < 1.0 were identified.
AOL 80mer FASTA3	• No matches were identified with >35% identity in any 80 AA match.
AOL Exact 8 AA	• No exact 8 AA matches were identified with AllergenOnline.
NCBI BLASTP Allergen	• No significant match was identified.
NCBI BLASTP "Toxin or Toxic" keyword search	• Modest scoring alignments were identified using toxin and toxic keywords to alcohol dehydrogenase enzymes of bacteria that are associated with toxicity (<i>C. raciborskii</i> , <i>P. agardhii</i> , <i>B. thuringiensis</i> , <i>B. cereus</i> , <i>E. coli</i> and <i>Streptococcus</i> sp.).
NCBI BLASTP No keyword	• The E-scores were small 2e-161 to 1e-140 and percent identities about 38–41%. No published evidence was found that these proteins are toxins.
Conclusion: it is unlikely that there is any risk for allergenicity or toxicity for this protein. Additionally, there are very good alignments to homologous proteins from many molds including <i>A. oryzae</i> and <i>Penicillium nalgiovense</i> . Thus, it is very unlikely that this is an allergen or a toxin.	

by Pariza and Foster (1983), Pariza and Johnson (2001) and several expert groups [(Berkowitz and Maryanski, 1989), (Scientific Committee f, 1992), (ouncil and Regul., 1990), (Jonas et al., 1996), (OECD, 1993), and (FAO/WHO, 1996)].

To address the potential allergenicity and toxicity of LegH Prep 2.0, a weight-of-evidence approach was employed. This multi-step assessment included performing mass spectrometry-based analysis on the protein and its heme cofactor. Following the CODEX 2003/2009 model for GM foods and novel food ingredients, bioinformatics, literature searches and

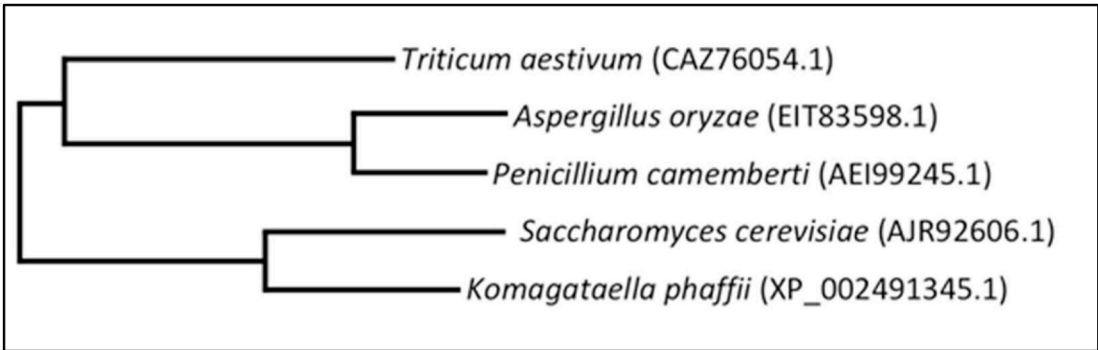


Fig. 7. Phylogenetic tree compares host *P. pastoris* protein GAPDH, isozyme 3 (XP_002491345.1) to sequences of matched allergen protein CAZ76054.1 of *T. aestivum*, and sequences of homologous proteins from edible fungi *A. oryzae*, *P. camemberti*, and *S. cerevisiae*. The edible fungal proteins were identified by BLASTP of the NCBI non-redundant database.

a pepsin digestion assay were performed to demonstrate the safety of LegH Prep 2.0 and its associated proteins. Currently, there isn't a single factor that is recognized as a sole identifier for protein allergenicity. Therefore, a weight-of-evidence approach, which takes into account a variety of factors and approaches for an overall assessment of allergenic potential was conducted based on the most current allergenicity assessment guidelines developed by the Codex (Alimentarius, 2009; Ladics et al., 2011).

IPMS analysis revealed that measurement of the quality attributes of the LegH by mass spectrometry confirms the biochemical identity of LegH from LegH Prep 2.0. In addition, a quantitative shotgun proteomics experiment was also performed on the LegH Prep 2.0 from three independent production lots which identified eleven *P. pastoris* proteins at $\geq 1\%$ of the total protein fraction, four of which have been previously analyzed (Jin et al., 2018; U.S. Food and Drug Administration, 2018), and seven of which are reported in this paper (Table 4). Bioinformatics analysis on the residual *Pichia* proteins in the LegH Prep. 2.0 resulted in

one significant match to the highly evolutionarily conserved GAPDH protein from wheat (*T. aestivum*). Results from the sequence homology analysis showed alignments to proteins identified as being from toxic organisms, but only as common homologues. GAPDH is a widely spread conserved protein whose homolog can be also found in many yeasts including *A. oryzae* and *P. roqueforti*, both which are widely used in the food industry, and from human.

Literature searches using PubMed, performed in August 2020 using the terms “allergen” AND “GAPDH” AND “wheat” revealed two publications. One of them was a publication by Sander et al., 2011 (Sander et al., 2011) which listed *T. aestivum* (*Tri a 34.0101*) as a putative allergen. In Sander's study, 19 *T. aestivum* proteins were heterologously expressed in *E. coli* and each was tested against sera samples from 40 bakers with work-related asthma/rhinitis (Sander et al., 2011). Only two sera samples out of 40 bound to the *T. aestivum* GAPDH expressed from *E. coli*, whereas 15 bakers had IgE to α -amylase inhibitors and 12 to at least one of the two cross-reactive carbohydrate determinants (CCDs)

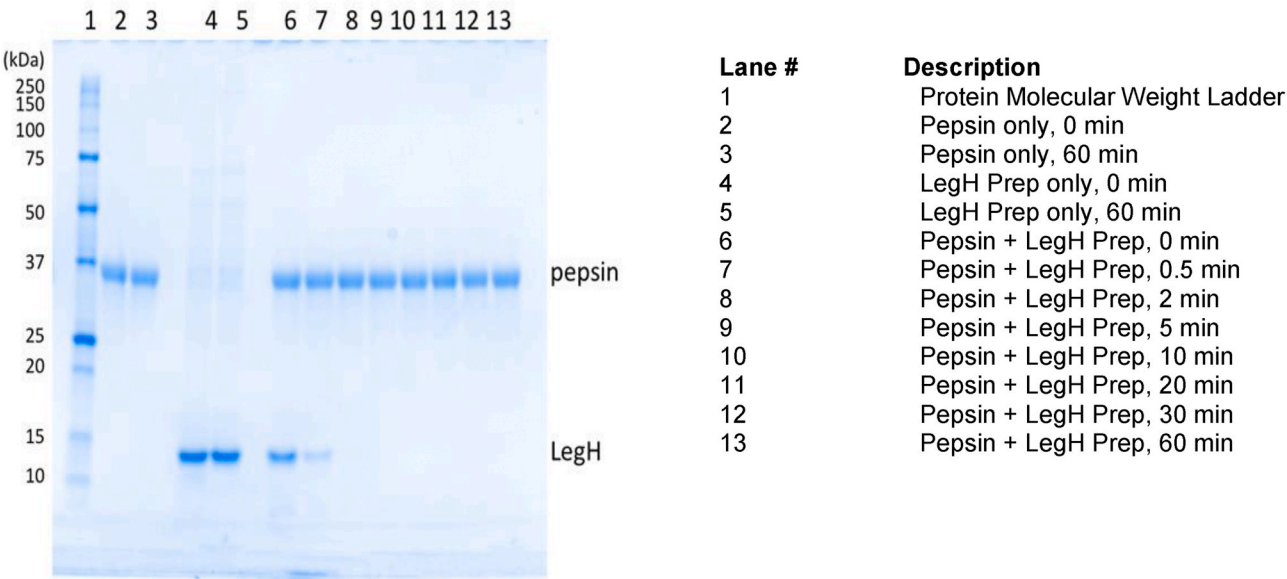


Fig. 8. Pepsin digestion of LegH final product analyzed on Bio-Rad Any kD Criterion TGX Protein Gel.

(Sander et al., 2011). A review of the other proteins in the test mix showed that most are likely more important allergens (e.g., CCDs at 80%, profilin at 60% and both the alpha-amylase inhibitors (38%) and thio-redoxin (30%)). The authors mentioned that peroxidase, serpin, dehydrin, LTP and omega-5 gliadin were very low. The authors concluded that the common proteins that are well known for allergy showed the most IgE binding in this study and a few proteins that are rarely implicated, like GAPDH, did not. Unfortunately, there is not enough data in this study that ties directly to a clinical response or other IgE to other proteins and the results presented do not indicate a common case of specific IgE to the GAPDH of wheat. GAPDH is a highly conserved protein that is found from *Archaea* to humans. Given the very low correlation of antigen binding and widely distribution of similar proteins, it is highly unlikely *T. aestivum* GAPDH is the causative agent of baker's asthma/rhinitis.

The second paper found in the literature search was a study by Zhu et al. (2020) where the authors analyzed the potential proteins in wheat that might be responsible for "wheat-dependent exercise-induced anaphylaxis" (WDEIA) in a Chinese population. As noted, WDEIA is a rare manifestation of food allergy and the underlying pathophysiological mechanisms of WDEIA are still not well understood. Other cofactors that have been identified as part of the onset of WDEIA, are acetylsalicylic acid and other non-steroidal anti-inflammatory drugs, alcohol, and infections (Scherf et al., 2015). WDEIA is thought to be caused by an unusual late-phase response in some subjects when they exercise within 4 h after consuming wheat. The authors undertook an extensive protein-IgE binding study with 33 subjects who appear to have WDEIA. In this study, Zhu et al. confirmed that omega-5 gliadin was the major allergen contributor in WDEIA and found that GAPDH tended to be more allergenic for most patients who had experienced WDEIA for the first time. The tests performed in the study showed that several GAPDH matched peptides might be the key segments related to allergy. However, few studies have actually confirmed the exact correlations between GAPDH and WDEIA. Hence, the mechanism and epitopes of GAPDH in WDEIA remain to be further studied (Zhu et al., 2020). This result suggests that IgE to GAPDH may have some relevance to severe reactions. However, a number of other proteins showed as strong of a correlation with severity, just not as frequent. Omega-5 gliadin is a well-studied food allergen and is one most researchers have found as likely causal proteins. Overall, results from the literature searches did not demonstrate a direct, common correlation between GAPDH and food allergy.

As part of the weight-of-evidence approach, soy LegH and its associated *Pichia* proteins were also subjected to a pepsin digest study. The LegH and its associated *Pichia* proteins were rapidly digested in SGF (pepsin) (Fig. 8). These results strongly suggest that no protease resistant protein is present in LegH Prep 2.0. While these results do not prove that risks of allergy are nonexistent, they demonstrate a very low likelihood of food allergy, or a need for serum IgE tests (Alimentarius, 2009; Ladics, 2008; Ladics et al., 2008). Also, a key characteristic that we must emphasize is that the average abundance of GAPDH in LegH Prep 2.0 is approximately 0.06%, which is quite low, far below most of the clinically important food allergens. Thus, the likely risk of food toxicity is low, considering that the FDA's "Approaches to Establish Thresholds for Major Food Allergens and for Gluten in Food" Guidelines [55] suggest that the minimum reported LOAEL observed for wheat allergen was 52.8 mg protein for allergic adults. For comparison, 100g of LegH Prep 2.0 would yield 60 mg GAPDH and the use levels of LegH Prep 2.0 would be far below that amount in food products. Based on all of the data

presented, as well as the very low amount of GAPDH in LegH Prep 2.0, there is a strong case to not perform serum IgE binding tests. Furthermore, the associated LegH Prep. 2.0 *Pichia* proteins are not the inserted protein or the primary product, rather they are minor constituents that are from the host. In terms of safety of the production organism, many pharmaceutical products have been made using *Pichia*, some of which are typically injected. If there were strong allergens or toxins in the commonly used *Pichia* lines, there would be many reports of adverse effects, which there are not, and it is unlikely the organism would be allowed for production of recombinant therapeutic proteins. In addition to the bioinformatics and proteomic data, a study obtained through the FOIA, showed the results of a *Pichia* feeding study in pigs. Dried *Pichia* was fed to growing/finishing swine at over 30% of the diet; results showed that the diet does not result in adverse effects (Supplement) (U. S. Food Additive Petition (FAP) #2207, 1993). This study supports and bolsters the safety of *Pichia* in foods, as the digestive system of swine is similar to that of humans.

Collectively, the literature search, the bioinformatics data (sequence homology) and the pepsin digest assay show that consumption of GAPDH from *P. pastoris* is unlikely to result in a food allergenic reaction. Overall, the sequence-related putative allergens identified in this search were not potent, nor were any of them known to be allergenic when orally consumed. As the LegH protein only makes up to 0.8% in the final plant-based meat products the abundance of each *Pichia* protein is in trace amounts. Furthermore, no published evidence was found to indicate that these proteins are toxins. Based on results from the proteomic analysis none of these proteins pose a significant risk of allergenicity.

The plant-based food industry has grown exponentially over the past few years. As more food-based technologies emerge, both the industry and regulators must come together to adapt to new challenges. In essence, plant-based foods are nothing new, with the exception of novel and innovative molecules making headway towards improving the taste and aroma and thereby, the experience of meat eating. The goal of producing these novel and innovate molecules is to revolutionize the way humans consume protein in a sustainable and environmentally responsible manner and for the overall well-being of the planet, while ensuring the safety of consuming these new ingredients.

Funding body information

Impossible Foods Inc. is the only source of funding for this study. All authors are employees of Impossible Foods Inc. Impossible Foods Inc. had a role in the study design; in the collection, analysis, and interpretation of data; in the writing of the manuscript, and in the decision to submit the article for publication.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. The authors declare the following financial interests/personal relationships, which may be considered as potential competing interests: The author(s) declared potential conflicts of interest with respect to the research, authorship, and/or publication of this article: Teresa F. Reyes, Yiming Chen, Rachel Z. Fraser, Teresa Chan and Xin Li are employees of Impossible Foods Inc.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.yrtph.2020.104817>.

References

- (AAFCO), 2013. 2013 Official Publication. American Association of Feed Control Officials Inc, USA.
- Alimentarius, Codex, 2003. Alinorm 03/34: Joint FAO/WHO Food Standard Programme, Codex Alimentarius Commission, Twenty-Fifth Session, Rome, Italy, 30 June- 5 July, 2003.
- Alimentarius, Codex, 2009. Foods derived from modern biotechnology, Annex 1, assessment of possible allergenicity, joint FAO/WHO food standard programme, codex Alimentarius commission, Rome, Italy. <http://www.fao.org/docrep/011/a1554e/a1554e00.htm> (Last assessed on May 4, 2020).
- Appleby, C.A., 1984. Leghemoglobin and Rhizobium respiration. *Annu. Rev. Plant Physiol.* 35, 443–478.
- Appleby, C.A., 1992. The origin and functions of haemoglobin in plants. *Sci. Prog.* 76, 365–398, 1933.
- Berkowitz, D., Maryanski, J., 1989. Codex Alimentarius Commission. Sess. vol. 18. Geneva, Switzerland, July 3–14, 1989.
- Cereghino, J.L., Cregg, J.M., 2000. Heterologous protein expression in the methylotrophic yeast *Pichia pastoris*. *FEMS Microbiol. Rev.* 1, 45–66.
- Chambers, M.C., MacLean, B., Burke, R., Amode, D., Ruderman, D.L., Neumann, S., Gatto, L., Fischer, B., Pratt, B., Egerton, J., Hoff, K., Kessner, D., Tasman, N., Shulman, N., Frewen, B., Baker, T.A., Brusniak, M.-Y., Paulse, C., Creasy, D., Flashner, L., Kani, K., Moulding, C., Seymour, S.L., Nuwaysir, L.M., Lefebvre, B., Kuhlmann, F., Roark, J., Rainer, P., Detlev, S., Hemenway, T., Huhmer, A., Langridge, J., Connolly, B., Chadick, T., Holly, K., Eckels, J., Deutsch, E.W., Moritz, R.L., Katz, J.E., Agus, D.B., MacCoss, M., Tabb, D.L., Mallick, P., 2012. A cross-platform toolkit for mass spectrometry and proteomics. *Nat. Biotechnol.* 10, 918–920. <https://doi.org/10.1038/nbt.2377>.
- Ciofalo, V., Barton, N., Kreps, J., Coats, I., Shanahan, D., 2006. Safety evaluation of a lipase enzyme preparation, expressed in *Pichia pastoris*, intended for use in the degumming of edible vegetable oil. *Regul. Toxicol. Pharmacol.* 1, 1–8.
- Colomer, M.S., Chailyan, A., Fennessy, R.T., Friis Olsson, K., Johnsen, L., Solodovnikova, N., Forster, J., 2020. Assessing population diversity of *Brettanomyces* Yeast Species and Identification of strains for brewing applications. *Front. Microbiol.* 11, 637.
- Di Girolamo, F., Muraca, M., Mazzina, O., Lante, I., Dahdah, L., 2015. Proteomic applications in food allergy, food allergenomics. *Food Allergy* 3, 259–266.
- FAO/WHO, 1996. Joint FAO/WHO Expert Consultation on Biotechnology and Food Safety. FAO/WHO, Rome, Italy.
- Federal Register, the Daily Journal of the United States Government, 2018. Color additive petition (CAP). <https://www.federalregister.gov/documents/2018/12/13/2018-26949/impossible-foods-inc-filing-of-color-additive-petition> (last accessed April 2020).
- Fraser, R., Brown, P.O., Karr, J., Holz-Schietinger, C., Cohn, E., 2017. Methods and compositions for affecting the flavor and aroma profile of consumables. U.S. Patent 14/797, 006.
- Fraser, R.Z., Shitut, M., Agrawal, P., Mendes, O., Klapholz, S., 2018. Safety evaluation of soy leghemoglobin protein preparation derived from *Pichia pastoris*, intended for use as a flavor catalyst in plant-based meat. *Int. J. Toxicol.* 3, 241–262.
- Galli, B.D., Martin, J.G.P., da Silva, P.P.M., Porto, E., Spoto, M.H.F., 2016. Sensory quality of Camembert-type cheese: Relationship between starter cultures and ripening molds. *Int. J. Food Microbiol.* 234, 71–75.
- Goldstein, B., Moses, R., Sammons, N., Birkved, M., 2017. Potential to curb the environmental burdens of American beef consumption using novel plant-based beef substitute. *PLoS One* 12 (12), e0189029.
- Goodman, R.E., 2008. Performing IgE serum testing due to bioinformatics matches in the allergenicity assessment of GM crops. *Food Chem. Toxicol.* 46 (Suppl. 10), S24–S34. <https://doi.org/10.1016/j.fct.2008.07.023>.
- Goodman, R.E., Ebisawa, M., Ferreira, F., Sampson, H.A., van Ree, R., Vieths, S., Baumert, J.L., Bohle, B., Lalithambika, S., Wise, J., Taylor, S.L., 2016. AllergenOnline: a peer-reviewed, curated allergen database to assess novel food proteins for potential cross-reactivity. *Mol. Nutr. Food Res.* 60, 1183–1198.
- Hardison, R., 1998. Hemoglobins from bacteria to man: evolution of different patterns of gene expression. *J. Exp. Biol.* 201, 1099–1117.
- Hardison, R.C., 2012. Evolution of hemoglobin and its genes. *Cold Spring Harb. Perspect. Med.* 2 (12), a011627.
- Hooda, J., Shah, A., Zhang, L., 2014. Heme, an essential nutrient from dietary proteins, critically impacts diverse physiological and pathological processes. *Nutrients* 3, 1080–1102.
- Jin, Y., He, X., Andoh-Kumi, K., Fraser, R.Z., Lu, M., Goodman, R.E., 2018. Evaluating potential risks of food allergy and toxicity of soy leghemoglobin expressed in *Pichia pastoris*. *Mol. Nutr. Food Res.* 62 (1–13), 1700297.
- Jonas, D.A., Antignac, E., Antoine, J.-M., Classen, H.-G., Huggett, A., Knudsen, I., Mahler, J., Ockhuizen, T., Smith, M., Teuber, M., 1996. The safety assessment of novel foods. Guidelines prepared by ILSI Europe Novel Food Task Force. *Food Chem. Toxicol.* 10, 931–940.
- Kitamoto, K., 2015. Cell biology of the Koji mold *Aspergillus oryzae*. *Biosci. Biotechnol. Biochem.* 79, 863–869.
- Kou, Q., Xun, L., Liu, X., 2016. TopPIC: a software tool for top-down mass spectrometry-based proteoform identification and characterization. *Bioinformatics* 32, 3495–3497.
- Krause, W., Ludwig, E., Rademacher, I., 1986. Enzymatic and chemical modification of animal blood. *Mol. Nutr.* 30, 299–302.
- Kreger-van Rijj, N.J.W., 1984. The Yeasts: a Taxonomic Study. Elsevier Science Publishers, Amsterdam.
- Küberl, A., Schneider, J., Thallinger, G.G., Anderl, I., Wibberg, D., Hajek, T., Jaenicke, S., Brinkrolf, K., Goessmann, A., Szczepanowski, R., Puhler, A., Schwab, H., Glieder, A., Pichler, H., 2011. High-quality genome sequence of *Pichia pastoris* CBS7435. *J. Biotechnol.* 4, 312–320.
- Ladics, G.S., 2008. Current codex guidelines for assessment of potential protein allergenicity. *Food Chem. Toxicol.* 46 (Suppl. 10), S20–S23.
- Ladics, G., Van Bilsen, J., Brouwer, H., Vogel, L., Vieths, S., Knippels, L.M.J., 2008. Assessment of three human FcεRI-transfected RBL cell-lines for identifying IgE induced degranulation utilizing peanut-allergic patient sera and peanut protein extract. *Regul. Toxicol. Pharmacol.* 51, 288–294. <https://doi.org/10.1016/j.yrtph.2008.04.012>.
- Ladics, G.S., Cressman, R.F., Herouet-Guicheney, C., Herman, R.A., Privalle, L., Song, P., Ward, J.M., McClain, S., 2011. Bioinformatics and the allergy assessment of agricultural biotechnology products: industry practices and recommendations. *Regul. Toxicol. Pharmacol.* 1, 46–53.
- Lin Cereghino, G.P., Sunga, A.J., Lin, C.J., Cregg, J.M., 2001. Expression of foreign genes in the yeast *Pichia pastoris*. *Genet. Eng. (N.Y.)* 23, 157–169.
- Mathesius, C.A., Barnett Jr., J.F., Cressman, R.F., Ding, J., Carpenter, C., Ladics, G.S., Schmidt, R.J., Layton, J., Zhang, X.Q., Appenzeller, L.M., Carlson, G., Ballou, S., Delaney, B., 2009. Safety assessment of modified acetocholactate synthase protein (GM-HRA) used as a selectable marker in genetically modified soybeans. *Regul. Toxicol. Pharmacol.* 3, 309–320. <https://doi.org/10.1016/j.yrtph.2009.08.003>.
- OECD, 1992. Safety Considerations for Biotechnology. Paris, France, p. 45.
- OECD, 1993. Safety Evaluation of Foods Derived by Modern Biotechnology: Concepts and Principles. Paris, France.
- Ofori-Anti, A.O., Ariyaratna, H., Chen, L., Lee, H.L., Pramod, S.N., Goodman, R.E., 2008. Establishing objective detection limits for the pepsin digestion assay used in the assessment of genetically modified foods. *Regul. Toxicol. Pharmacol.* 2, 94–103.
- I. F. B. Council. *Regul. Toxicol. Pharmacol.* 12, 1990, S114.
- Pariza, M., Foster, E.J., 1983. Determining the safety of enzymes in food processing. *Food Prot.* 46, 453.
- Pariza, M.W., Johnson, E.A., 2001. Evaluating the safety of microbial enzyme preparations in food process: an update for a new century. *Regul. Toxicol. Pharmacol.* 33, 173.
- Pharmacopeia, U.S., 2018. Enzyme preparations. In: Food Chemicals Codex, eleventh ed., pp. 413–419 Washington, D.C.
- Sander, I., Rozynek, P., Rihs, H.P., van Kampen, V., Chew, F.T., Lee, W.S., Kotschy-Lang, N., Merget, R., Brüning, T., Raulf-Heimsoth, M., 2011. Multiple wheat flour allergens and cross-reactive carbohydrate determinants bind IgE in baker's asthma. *Allergy* 66, 1208–1215.
- Scherf, K.A., Brockow, K., Biedermann, T., Koehler, P., Wieser, H., 2015. Wheat-dependent exercise-induced anaphylaxis. *Clin. Exp. Allergy* 46 (1), 10–20.
- EU Scientific Committee for Food, European Commission, Luxembourg, 1992, p. 13.
- Siruguri, V., Bharatraj, D.K., Vankudavath, R.N., Mendu, V.V., Gupta, V., Goodman, R.E., 2015. Evaluation of Bar, Barnase, and Barstar recombinant proteins expressed in genetically engineered *Brassica juncea* (Indian mustard) for potential risks of food allergy using bioinformatics and literature searches. *Food Chem. Toxicol.* 83, 93–102.
- Thomas, K., Aalbers, M., Bannon, G.A., Bartels, M., Dearman, R.J., Esdaile, D.J., Fu, T.J., Glatt, C.M., Hadfield, N., Hatzos, C., Hefle, S.L., Heylings, J.R., Goodman, R.E., Henry, B., Herouet, C., Hosapple, M., Ladics, G.S., Landry, T.D., MacIntosh, S.C., Rice, E.A., Privalle, L.S., Steiner, H.Y., Teshima, R., Van Ree, R., Woolhiser, M., Zawodny, J., 2004. A multi-laboratory evaluation of a common in vitro pepsin digestion assay protocol used in assessing the safety of novel proteins. *Regul. Toxicol. Pharmacol.* 2, 87–89.
- Thompson, C.A., 2010. FDA Approves kallikrein inhibitor to treat hereditary angioedema. *Am. J. Health Syst. Pharm.* 2, 93. <https://doi.org/10.2146/news100005>.
- Tyanova, S., Temu, T., Cox, J., 2016. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nat. Protoc.* 11, 2301–2319.

- U.S. Food Additive Petition (FAP) #2207, 1993. Food additives permitted in feed and drinking water of animals; *Pichia pastoris* dried yeast. FR Doc. 93, 27392.
- U.S. Food and Drug Administration, 2006. Center for food safety and applied nutrition, office of food additive safety. Agency Response Letter GRAS Notice No. GRN 000204. <https://wayback.archiveit.org/7993/20171031020303/https://www.fda.gov/Food/IngredientsPackagingLabeling/GRAS/NoticeInventory/ucm153734.htm> (Last accessed August 2020).
- U.S. Food and Drug Administration, 2018. Center for food safety and applied nutrition, office of food additive safety. Agency Response Letter GRAS Notice No. GRN 000737. <https://www.fda.gov/media/116243/download> (Last accessed November, 2019).
- Zhang, Y., Fonslow, B.R., Shan, B., Baek, M.C., Yates III., J.R., 2013. Protein analysis by shotgun/bottom-up proteomics. *Chem. Rev.* 4, 2343–2394.

- Zhu, Y.Q., Wang, D.Q., Liu, B., Hu, Y., Shen, Y.Y., Xu, J.H., Tang, H., 2020. Wheat-dependent exercise-induced anaphylaxis in Chinese people: a clinical research on 33 cases with antigenic analysis of wheat proteins. *Clin. Exp. Dermatol.* 45, pp56–62. <https://doi.org/10.1111/ced.14035>.

further reading

- Gonzalez, L.M., Moeser, A.J., Blikslager, A.T., 2015. Porcine models of digestive disease: the future of large animal translational research. *Transl. Res.* 1, 12–27.

Supplemental Material

Swine Feeding Study obtained through FOIA request

The swine feeding study utilized crossbred commercial pigs (N=64) that were evenly distributed into four groups (US FDA FAP, 1993). Each group was administered *P. pastoris* to replace the soybean meal typically formulated in the diet, thereby replacing 0, 35%, 65% or 100% of the soybean meal (on a weight/weight basis). Based on publicly available information on diet formulations indicating a corn: soybean meal contains approximately 23.4% soybean meal (National Research Council, 1998), the amount of *P. pastoris* SCP administered was calculated at 0, 8.9%, 15.2% or 31.4% of the diet (US FDA, FAP 1993). At the end of the study, body weights were obtained from a subset of pigs from each group, blood was collected for hematology and clinical chemistry parameters, urine pH analysis, specific gravity and urine protein end point analysis.

Analysis of red or white blood cell counts, blood urea nitrogen (BUN), packed cell volume (PCV), alkaline phosphatase, serum aspartate aminotransferase (AST) and alanine aminotransferase (ALT) did not reveal any significant effects that were outside normal limits or were of a consistent, toxicological nature (US FDA, FAP 2018) (Supplemental Table S1). In the same fashion, there were no changes to urinalysis or serum biochemistry parameters that were treatment-related or toxicological in nature (US FDA, FAP 1993). A gross evaluation of the liver, gastrointestinal tract, spleen, kidney and lung found no adverse effects from the consumption of the *P. pastoris* ingredient (US FDA, FAP 1993). Histopathological analysis of the liver, gastrointestinal tract, kidney and lymph was also negative for any adverse effect (US FDA, FAP 1993). The majority of the mean organ and carcass weights of the pigs fed *P. pastoris* were not significantly different from the control diet ($P>0.10$), except for a trend of significance ($P<0.25$) for a non-dose-dependent increase in the small intestine weight in the low dose group, and a trend ($P<0.10$) in decreasing cecum weight (US FDA, FAP 1993), (data not shown). Overall, the data

obtained from this study show that consumption of dried *P. pastoris* by growing/finishing swine at over 30% of the diet does not result in adverse effects. This study supports the safety of *P. pastoris* in foods, as the digestive system of swine is similar to that of humans (Gonzalez, 2015).

Table S1. Serum chemistry, hematology and urinalysis evaluation of swine consuming *P. pastoris* in the diet.

	<i>P. pastoris</i> SCP (% of diet)				%CV
End points#	0	8.9	15.2	31.4	
Body weight (kg)	87	92.3	93.0*	86.8	7.8
Clinical Chemistry					
Alkaline phosphatase (sigma)**	4.0	5.0	3.5	3.1	35.6
BUN (mg/dl)	17.8	16.3	20.4	19.1	22.5
AST (SF/ml)**	29	36	31	32	38.0
ALT (SF/ml)**	19	33	22	22	35.6+
Protein (g/dl)	9.8	9.1	10.2	9.7	30.0
Hematology					
RBC (X 10⁶/ml)	8.43	8.28	7.80	7.89	7.2
WBC (X10³/ml)	22.1	20.9	23.6	19.1	11.0+
Hb (g/dL)	13.5	14.1	13.8	13.0	9.0
PCV (%)	38.2	36.2	36.8	37.9	21.1
Urinalysis					
Specific gravity	1.024	1.021	1.023	1.020	
Urinary protein	±	±	±	±	
pH	6.1	6.5	6.2	6.0	

#Tabular format of the results indicated four pigs analyzed/group, while the methodology (hematology/clinical chemistry) statements indicated blood was obtained from two pigs/group; urinalysis was obtained from N=2 pigs/sex/group.
+P< 0.25;
*Average body weight for this group was stated as both 93.0 and 93.1 kg; for safety assessments the lower body weight was utilized.
**ALT (alanine aminotransferase) and AST (aspartate aminotransferase) enzyme values determined by the Sigma Frankel method (Kaneko and Cornelius, 1970), Alkaline phosphatase values determined by the Sigma method.
Alk P=Alkaline phosphatase; BUN=blood urea nitrogen; CV=Coefficient of variation; Hb=hemoglobin; PCV=packed cell volume; RBC=red blood cells; SCP=Single Cell Protein; WBC=white blood cells.

Table S2. Peptides from LegH identified in the Shotgun Proteomics experiment

Sequence	Missed cleavages	Mass	Start position	End position	Score
GAFTEKQEALVSSSF EAFK	1	2075.021	2	20	241.89
QEALVSSSF EAFK	0	1441.709	8	20	237.84
ANIPQYSVVFYTSILE K	0	1971.035	21	37	203.85
ANIPQYSVVFYTSILE KAPAAK	1	2409.294	21	42	120.75
APAAKDLFSFLSNGV DPSNPK	1	2174.101	38	58	218.63
DLFSFLSNGVDPSN PK	0	1735.842	43	58	278.93
DLFSFLSNGVDPSN PKLTGHA EK	1	2472.229	43	65	211.75
LTGHA EKLFGLVR	1	1439.825	59	71	53.034
DSAGQLK	0	717.3657	72	78	149.4
DSAGQLKANGTVVA DAALGSIHAQK	1	2421.261	72	96	225.11
ANGTVVADAALGSIH AQK	0	1721.906	79	96	367.85
ANGTVVADAALGSIH AQKAITDPQFVVVK	1	2919.582	79	107	116.6
AITDPQFVVVK	0	1215.686	97	107	209.21
AITDPQFVVVKEALL K	1	1770.029	97	112	198.23
EALLKTIK	1	914.5801	108	115	174.27

TIKEAVGDK	1	959.5288	113	121	88.364
TIKEAVGDKWSDELS SAWEVAYDELAAAIK	2	3294.63	113	142	106.39
EAVGDKWSDELSSA WEVAYDELAAAIK	1	2952.403	116	142	447.9
EAVGDKWSDELSSA WEVAYDELAAAIKK	2	3080.498	116	143	271.95
WSDELSSAWEVAYD ELAAAIK	0	2353.111	122	142	320.35
WSDELSSAWEVAYD ELAAAIKK	1	2481.206	122	143	298.78

Table S3. Detailed precursor information of the major proteoforms for LegH isoforms from soy root nodule.

Isoform	Precursor m/z	Precursor charge	Precursor mass	Matched peaks	Matched fragments	p-value
LegH A	1172.7039	13+	15232.056	109	83	9.85e-63
LegH C1	1271.5896	12+	15246.987	85	74	5.53e-56
LegH C2	1283.0095	12+	15384.027	68	63	3.70e-50
LegH C3	1287.7672	12+	15441.122	93	83	2.52e-63

Table S4. Matched mass peaks for major proteoform of LegH A.

Peak	Observed Mono mass	Observed Mono m/z	Intensity	Charge	Theoretical mass	Ion	Mass error (ppm)
1	2383.2177	1192.6161	8103196	2	2383.206	B22	4.9
2	10034.4406	1115.9451	5739584	9	10034.37	Y96	7.01
3	9578.2423	1065.2564	5088465	9	9578.173	Y91	7.2
4	9963.4103	1108.0529	4881494	9	9963.333	Y95	7.75
5	5654.9223	1414.7378	4455846	4	5654.887	B52	6.33
6	10033.4199	1255.1848	3781624	8	10034.37	Y96	5.19

7	9963.3838	1246.4303	3847513	8	9963.333	Y95	5.09
8	10146.5304	1128.3995	3104211	9	10147.45	Y97	7.73
9	9849.3679	1095.3815	2607272	9	9849.29	Y94	7.89
10	4109.1262	1370.716	3593219	3	4109.104	B37	5.46
11	9577.2047	1198.1579	2882107	8	9578.173	Y91	3.51
12	9693.2657	1212.6655	2639966	8	9693.2	Y92	6.74
13	9848.3223	1232.0476	2479282	8	9849.29	Y94	3.5
14	4856.6698	972.3412	2388285	5	4856.648	Y44	4.46
15	12849.9495	1169.1845	3438379	11	12849.85	Y121	7.44
16	10380.6351	1154.4112	2120331	9	10381.55	Y99	7.96
17	6490.5443	1082.7647	1828012	6	6490.491	Y61	8.29
18	10294.6038	1144.8521	1822238	9	10294.52	Y98	7.87
19	10376.4719	1298.0663	2521869	8	10376.41	B99	5.79
20	2270.1435	1136.079	1958692	2	2270.122	B21	9.49
21	10146.4961	1269.3193	1785678	8	10147.45	Y97	4.35
22	12274.6382	1228.4711	1849596	10	12275.58	Y116	5.03
23	6391.4519	1279.2977	1824725	5	6391.422	Y60	4.67
24	12848.953	1285.9026	2812883	10	12849.85	Y121	7.9
25	1102.5769	1103.5842	1309684	1	1102.566	B10	10.05
26	5198.7079	1300.6843	2011860	4	5198.69	B47	3.52
27	4348.2659	1450.4292	1081882	3	4348.231	B40	8.09
28	8742.6372	1249.9555	1773570	7	8742.569	B82	7.76
29	9693.2764	1078.038	933854.6	9	9693.2	Y92	7.85
30	5950.2093	1191.0491	1820870	5	5950.2	Y55	1.61
31	8911.7326	1274.1119	756811.3	7	8912.675	B84	6.73
32	5269.7476	1318.4442	1598523	4	5269.727	B48	3.95
33	6490.5188	1299.111	2035562	5	6490.491	Y61	4.37

34	5383.8252	1346.9636	1720828	4	5383.77	B49	10.31
35	6319.4646	1054.2514	934310.6	6	6320.385	Y59	12.98
36	6391.4763	1066.2533	1143492	6	6391.422	Y60	8.49
37	9791.3297	1088.9328	949958.7	9	9792.269	Y93	6.47
38	12175.5879	1218.5661	1548967	10	12176.51	Y115	6.56
39	1201.6423	601.8284	611414.9	2	1201.634	B11	6.68
40	989.4866	990.4939	744528.4	1	989.4818	B9	4.86
41	10294.5305	1287.8236	912606.7	8	10294.52	Y98	0.76
42	4187.2939	1047.8308	634084.1	4	4187.263	Y38	7.37
43	8398.447	1200.7854	797453.6	7	8398.4	B78	5.62
44	6905.7527	1151.9661	922114.9	6	6905.697	Y66	8.04
45	1201.6429	1202.6501	978970.1	1	1201.634	B11	7.14
46	2156.0888	1079.0517	754780.2	2	2156.079	B20	4.53
47	4059.1687	1015.7995	583957	4	4059.168	Y37	0.14
48	5706.0937	1142.226	678452.2	5	5707.078	Y52	3.2
49	3056.5853	1529.2999	759417.3	2	3056.55	B28	11.72
50	11865.4379	1187.5511	872787.2	10	11866.38	Y113	5.19
51	1102.5708	552.2927	596661.4	2	1102.566	B10	4.44
52	9791.3257	1224.923	922562.7	8	9792.269	Y93	6.05
53	8198.3029	1367.3911	926195.8	6	8199.268	B76	4.58
54	4286.3774	1072.6016	774232.4	4	4286.332	Y39	10.71
55	11323.1629	1133.3236	500613.5	10	11323.09	Y108	6.6
56	2233.196	745.4059	510080.3	3	2233.174	Y20	9.7
57	9265.1017	1159.145	1301088	8	9266.03	Y88	7.99
58	5950.2647	992.7181	655126.8	6	5950.2	Y55	10.93
59	2383.2128	795.4116	435566.1	3	2383.206	B22	2.88
60	5707.1146	952.193	398005.1	6	5707.078	Y52	6.45

61	3056.5662	1019.8627	619199.9	3	3056.55	B28	5.45
62	2608.3458	1305.1802	756714.2	2	2608.317	B24	10.93
63	3162.6445	1055.2221	661880.9	3	3162.599	Y29	14.5
64	6320.4675	1265.1008	1350166	5	6320.385	Y59	13.07
65	6064.3201	1213.8713	1211174	5	6063.284	Y56	5.61
66	4277.1997	1426.7405	535402.7	3	4277.194	B39	1.43
67	2957.5197	1479.7671	570549.3	2	2957.481	B27	13.04
68	3504.8257	877.2137	274916.1	4	3504.825	Y32	0.12
69	11195.0355	1120.5108	470401.7	10	11194.99	Y107	3.78
70	1809.887	905.9508	328682.1	2	1809.879	B17	4.69
71	2921.4988	974.8402	331113.5	3	2921.456	Y26	14.65
72	5242.8623	1311.7229	718126.2	4	5242.828	Y48	6.5
73	10527.6633	1316.9652	608592.8	8	10528.62	Y100	4.03
74	2771.388	924.8033	446231.1	3	2771.381	B25	2.65
75	6746.678	1125.4536	725418.4	6	6747.628	Y64	7.76
76	6133.3039	1023.2246	411545	6	6134.321	Y57	-2.39
77	6062.3088	1011.3921	405167.3	6	6063.284	Y56	4.51
78	6588.5983	1099.107	479739.9	6	6589.559	Y62	6.34
79	2085.0543	1043.5344	397199.9	2	2085.042	B19	5.98
80	2749.4199	917.4806	306227.1	3	2749.408	Y24	4.5
81	9266.1311	1030.5774	425465.5	9	9266.03	Y88	10.91
82	1288.6654	645.34	358420.8	2	1288.666	B12	-0.73
83	671.4367	672.444	439828.3	1	671.433	Y7	5.52
84	918.4472	919.4545	264372	1	918.4447	B8	2.71
85	9026.7864	1129.3556	539542.2	8	9027.702	B85	9.63
86	4484.5216	1122.1377	537188.2	4	4484.468	Y41	11.88
87	4385.4129	878.0898	171933	5	4385.4	Y40	2.95

88	5498.9966	1100.8066	338339.3	5	5498.982	Y50	2.69
89	1738.8513	870.4329	313634.4	2	1738.841	B16	5.7
90	600.3999	601.4072	363398	1	600.3959	Y6	6.59
91	6834.6987	1140.1237	751248.3	6	6834.66	Y65	5.65
92	989.4855	495.75	248455	2	989.4818	B9	3.74
93	6205.3659	1242.0805	590549.9	5	6205.358	Y58	1.28
94	5085.5942	1272.4058	815682.7	4	5085.606	B46	-2.24
95	1956.9615	979.488	253857	2	1956.947	B18	7.43
96	2858.4184	953.8134	200675.2	3	2858.413	B26	2
97	4059.1958	1354.0726	455832	3	4059.168	Y37	6.83
98	6205.4504	1035.249	482677.1	6	6205.358	Y58	14.89
99	8086.2926	1348.7227	554120.7	6	8086.184	B75	13.47
100	3203.6416	1602.8281	355216.8	2	3203.618	B29	7.41
101	9097.7249	1138.2229	253545.3	8	9098.739	B86	-1.29
102	5371.9078	1075.3888	187736.1	5	5370.923	Y49	-3.3
103	3667.8804	1834.9475	208166.4	2	3667.845	B33	9.65
104	4484.455	897.8983	153947.4	5	4484.468	Y41	-2.96
105	10755.8435	1345.4877	241810.2	8	10756.73	Y102	10.38
106	2957.5114	986.8444	99161.77	3	2957.481	B27	10.24
107	3909.9891	1956.0018	139738.9	2	3909.972	B35	4.45
108	6904.6952	987.3923	118019	7	6905.697	Y66	0.06
109	2621.3096	874.7772	108887.8	3	2621.313	Y23	-1.13

Table S5. Matched mass peaks for major proteoform of LegH C1.

Peak	Observed Mono mass	Observed Mono m/z	Intensity	Charge	Theoretical mass	Ion	Mass error (ppm)
------	--------------------	-------------------	-----------	--------	------------------	-----	------------------

1	10063.4176	1258.9345	4795048	8	10064.33	Y96	8.62
2	9992.3809	1250.0549	3653532	8	9993.296	Y95	8.72
3	2355.194	1178.6043	5383866	2	2355.175	B22	8.2
4	9607.2042	1201.9078	3952735	8	9608.136	Y91	7.31
5	5639.8901	1410.9798	3504426	4	5639.85	B52	7.02
6	10410.5481	1302.3258	2259438	8	10411.52	Y99	3.15
7	10176.4423	1273.0626	2641627	8	10177.42	Y97	2.7
8	9878.3565	1235.8018	2104355	8	9879.253	Y94	10.7
9	10557.6492	1320.7134	1595237	8	10558.59	Y100	6.2
10	5254.7314	1314.6901	1495432	4	5254.691	B48	7.75
11	10473.4905	1310.1936	1147666	8	10474.46	B99	3.15
12	10323.529	1291.4484	1721092	8	10324.49	Y98	4.42
13	4094.0859	1365.7026	2535387	3	4094.068	B37	4.44
14	2780.3358	1391.1752	2226631	2	2780.329	Y26	2.32
15	4773.5469	1194.394	1350171	4	4773.527	Y44	4.19
16	6463.4529	1293.6979	1251095	5	6463.432	Y61	3.25
17	9722.2383	1216.2871	773758.9	8	9723.163	Y92	7.96
18	5368.7724	1343.2004	1356901	4	5368.734	B49	7.23
19	4689.4405	1173.3674	1259881	4	4689.401	B43	8.5
20	2242.1005	1122.0575	796334.2	2	2242.091	B21	4.4
21	10064.3952	1438.778	540017.6	7	10064.33	Y96	6.16
22	12218.5656	1358.6257	973156.9	9	12219.47	Y115	8.13
23	6934.685	1387.9443	864139	5	6935.66	Y66	3.93
24	9296.0473	1329.014	580742.3	7	9295.993	Y88	5.84
25	2279.1486	1140.5816	546033.9	2	2279.132	Y21	7.2
26	9608.1982	1373.607	921423.4	7	9608.136	Y91	6.44
27	2128.0597	1065.0371	405762.2	2	2128.048	B20	5.65

28	9821.2436	1228.6627	682829.9	8	9822.232	Y93	1.45
29	9722.1823	1389.8905	715536.6	7	9723.163	Y92	2.2
30	4576.3439	1145.0932	567357.4	4	4576.317	B42	5.96
31	9993.3451	1428.628	699428.8	7	9993.296	Y95	4.91
32	10785.7429	1349.2251	855828.6	8	10786.7	Y102	4.46
33	3028.5372	1515.2759	621964.6	2	3028.518	B28	6.29
34	2879.4061	1440.7103	636020.9	2	2879.398	Y27	2.9
35	8312.4355	1188.4981	432320.4	7	8312.327	B77	13.09
36	2035.085	1018.5498	301096.5	2	2035.063	Y19	10.98
37	1074.5442	1075.5515	372997.7	1	1074.535	B10	9
38	3976.085	1326.3689	378549.5	3	3976.047	Y37	9.56
39	1922.0065	962.0105	520920.3	2	1921.979	Y18	14.52
40	8312.4288	1386.4121	511372.4	6	8312.327	B77	12.29
41	3079.4972	1540.7559	503259.2	2	3079.478	Y29	6.4
42	2164.1159	1083.0652	287063.4	2	2164.105	Y20	4.9
43	2743.369	915.4636	329031.9	3	2743.349	B25	7.17
44	9295.0749	1162.8916	333914.6	8	9295.993	Y88	9.06
45	4262.1932	1421.7383	463859.1	3	4262.158	B39	8.35
46	3028.5265	1010.5161	361027.6	3	3028.518	B28	2.73
47	4203.2468	1402.0895	346856.2	3	4203.21	Y39	8.65
48	1676.8874	839.451	215054.9	2	1676.877	Y15	5.97
49	3421.7242	1141.582	370254.8	3	3421.704	Y32	5.84
50	5524.8788	1382.227	591069	4	5524.824	B51	10
51	4836.5103	1210.1348	593542	4	4836.469	B44	8.52
52	9832.2155	1230.0342	451449	8	9832.126	B93	9.09
53	1191.6452	1192.6525	406940	1	1191.65	Y11	-4.01
54	8526.4678	1219.0741	259766.6	7	8527.417	B79	6.2

55	4333.2572	1445.4263	955730.6	3	4333.195	B40	14.41
56	1834.9667	918.4906	214942.3	2	1834.947	Y17	10.97
57	1074.5396	538.2771	180179.3	2	1074.535	B10	4.72
58	1781.8492	891.9319	227870.6	2	1781.847	B17	1.1
59	5288.8302	1323.2148	282818.6	4	5287.802	Y49	4.87
60	1173.6113	587.8129	208818.1	2	1173.603	B11	7.08
61	2355.182	786.0679	156611.7	3	2355.175	B22	3.09
62	8953.7326	1280.1119	400591.2	7	8954.66	B84	8.33
63	3539.736	1180.9193	384940.7	3	3539.725	B32	3.13
64	1710.8259	856.4202	157379	2	1710.81	B16	9.25
65	2929.4601	1465.7373	456587.9	2	2929.45	B27	3.52
66	4923.5727	1231.9005	290028.9	4	4923.501	B45	14.55
67	3207.6009	1070.2076	158244.8	3	3207.573	Y30	8.87
68	5624.0191	1125.8111	238720.8	5	5623.957	Y52	11.11
69	913.5645	457.7895	126436.8	2	913.5597	Y9	5.24
70	2666.2676	1334.1411	263206.7	2	2666.286	Y24	-7.06
71	6365.3838	1274.084	919961.7	5	6364.364	Y60	2.81
72	1747.9283	874.9714	98580.06	2	1747.915	Y16	7.89
73	1260.6439	631.3292	108460	2	1260.635	B12	7.04
74	2057.0187	1029.5166	247739.1	2	2057.011	B19	3.95
75	1490.7927	1491.7999	201757	1	1490.798	Y14	-3.65
76	3175.6208	1059.5475	150915.6	3	3175.587	B29	10.77
77	4023.0728	1342.0315	425461.6	3	4023.031	B36	10.5
78	1361.7608	681.8877	95861.91	2	1361.756	Y13	3.92
79	2830.4068	944.4762	99152.69	3	2830.381	B26	8.97
80	8783.5768	1255.804	975524.4	7	8784.555	B82	2.77
81	6935.7172	1156.9601	222472.1	6	6935.66	Y66	8.23

82	600.4003	601.4075	85798.51	1	600.3959	Y6	7.21
83	3320.676	1107.8993	108111.5	3	3320.657	Y31	5.86
84	3976.0776	995.0267	124991.9	4	3976.047	Y37	7.71
85	3338.6935	1113.9051	133851	3	3338.65	B30	13.05

Table S6. Matched mass peaks for major proteoform of LegH C2.

Peak	Observed Mono mass	Observed Mono m/z	Intensity	Charge	Theoretical mass	Ion	Mass error (ppm)
1	10213.4511	1277.6887	2178334	8	10213.38	Y97	6.88
2	9741.2407	1218.6624	1850160	8	9741.189	Y92	5.3
3	5642.8749	1411.726	1548244	4	5642.85	B52	4.4
4	2355.1886	1178.6016	1886621	2	2355.175	B22	5.92
5	10126.4174	1266.8094	1025135	8	10126.35	Y96	6.77
6	10325.4813	1291.6924	1163461	8	10326.46	Y98	1.81
7	10472.5471	1310.0757	754660.4	8	10473.53	Y99	1.54
8	10559.5842	1320.9553	1150251	8	10560.57	Y100	2
9	4081.0923	1361.3714	1067985	3	4081.072	B37	4.87
10	10011.3662	1252.4281	897547.3	8	10012.31	Y95	6.26
11	10707.6041	1339.4578	780456.6	8	10707.63	Y101	-2.77
12	4992.6729	1249.1755	548043.4	4	4992.653	Y45	4.01
13	9855.2692	1232.9159	352820	8	9856.216	Y93	5.63
14	4320.2277	1441.0832	426444.9	3	4320.199	B40	6.54
15	9741.2264	1392.611	342302.7	7	9741.189	Y92	3.83
16	10391.4319	1299.9363	353250.2	8	10391.39	B99	4.39
17	5257.7529	1315.4455	431308.7	4	5257.69	B48	11.9
18	7095.7612	1420.1595	334206.4	5	7096.744	Y67	2.72

19	5371.7401	1343.9423	458631.8	4	5371.733	B49	1.27
20	5171.6845	1293.9284	639860.5	4	5170.658	B47	4.6
21	6654.5271	1331.9127	707880.7	5	6654.527	Y62	0.08
22	8728.5748	1247.9465	479603.9	7	8729.513	B82	7.39
23	2242.0961	1122.0553	233658.4	2	2242.091	B21	2.44
24	8287.3818	1184.919	181143.8	7	8287.295	B77	10.47
25	4676.4264	1170.1139	390219.2	4	4676.405	B43	4.48
26	9954.334	1245.299	347016.8	8	9955.284	Y94	5.21
27	10935.7726	1367.9789	356825.1	8	10935.74	Y103	2.54
28	8827.6668	1262.1025	226994.5	7	8828.581	B83	9.98
29	4563.3424	1141.8429	214804.3	4	4563.321	B42	4.61
30	11744.2314	1305.9219	364563.9	9	11743.21	B111	1.7
31	11744.2314	1305.9219	364563.9	9	11744.23	Y111	0.5
32	6555.5017	1312.1076	429580.9	5	6555.458	Y61	6.64
33	2128.0563	1065.0354	161359.4	2	2128.048	B20	4.04
34	2311.1864	1156.6005	183494.7	2	2311.174	Y21	5.53
35	1338.7193	1339.7266	160971.4	1	1338.718	Y12	0.67
36	3028.5371	1010.5196	142903.4	3	3028.518	B28	6.24
37	1074.5376	1075.5449	111119.5	1	1074.535	B10	2.86
38	1173.6114	1174.6187	175885.9	1	1173.603	B11	7.15
39	5391.8546	1348.9709	321330.9	4	5392.849	Y49	1.52
40	2069.0648	1035.5397	80236.77	2	2069.047	Y19	8.62
41	4448.3028	1483.7749	152487.4	3	4448.294	B41	1.89
42	2827.4071	943.4763	110331.6	3	2827.407	Y25	0.08
43	3175.5867	1059.5362	86052.56	3	3175.587	B29	0.04
44	3439.7011	1147.5743	100874.6	3	3439.698	B31	1
45	11302.9861	1256.8946	334546.3	9	11302.97	Y107	1.71

46	6114.3	1223.8673	152765.6	5	6114.236	Y56	10.5
47	9268.8	1325.1216	187974.2	7	9269.803	B88	-0.11
48	5942.0093	1486.5096	112626.4	4	5940.978	B55	4.89
49	6753.6152	1351.7303	324079.3	5	6753.595	Y63	3
50	2743.3434	915.4551	87849.94	3	2743.349	B25	-2.17
51	6483.483	1297.7039	291304.9	5	6484.421	Y60	9.91
52	4195.1828	1399.4015	183543.1	3	4195.173	Y38	2.34
53	4910.5431	1228.6431	132954.4	4	4910.506	B45	7.59
54	4521.4311	1508.151	118869.7	3	4521.405	Y41	5.83
55	492.3082	493.3154	63308.26	1	492.3061	Y4	4.26
56	1173.6087	587.8116	49497.81	2	1173.603	B11	4.9
57	961.4444	962.4517	66345.39	1	961.4505	B9	-6.31
58	1074.5376	538.2761	48045.23	2	1074.535	B10	2.79
59	2355.2014	786.0744	52405.66	3	2355.175	B22	11.33
60	1060.6346	531.3246	55898.22	2	1060.628	Y10	6.13
61	7095.7901	1183.639	136782.9	6	7096.744	Y67	6.8
62	1781.838	891.9263	83630.36	2	1781.847	B17	-5.14
63	1928.9291	965.4718	72198.26	2	1928.916	B18	6.99
64	6056.2107	1212.2494	99301.01	5	6057.214	Y55	-0.22
65	1710.8236	856.4191	48245.42	2	1710.81	B16	7.89
66	5856.07	1172.2213	90381.36	5	5857.098	Y53	-4.42
67	761.374	762.3813	30615.06	1	761.3708	B7	4.21
68	4195.186	1049.8038	43414.77	4	4195.173	Y38	3.12
69	3298.62	1100.5473	36613.36	3	3298.603	Y30	5.04

Table S7. Matched mass peaks for major proteoform of LegH C3.

Peak	Mono mass	Mono m/z	Intensity	Charge	Theoretical	Ion	Mass
------	-----------	----------	-----------	--------	-------------	-----	------

					mass		error (ppm)
1	10225.5634	1279.2027	1694520.49	8	10226.4376	Y97	12.53
2	9769.3487	1222.1759	1411490.4	8	9770.2408	Y92	11.29
3	10154.5273	1270.3232	1368069.56	8	10155.4005	Y96	12.71
4	5670.9325	1418.7404	1169275.62	4	5670.8814	B52	9.01
5	10338.5833	1293.3302	910686.54	8	10339.5217	Y98	6.18
6	2371.193	1186.6038	1435065.28	2	2371.1696	B22	9.88
7	10573.6691	1322.7159	1071899.29	8	10573.6222	Y100	4.44
8	10041.4926	1256.1938	775843.88	8	10041.3576	Y95	13.44
9	6695.6195	1340.1312	830843.02	5	6696.5735	Y62	7.21
10	10485.6604	1311.7148	590380.01	8	10486.5901	Y99	6.92
11	4097.0914	1366.7044	757058.48	3	4097.0674	B37	5.86
12	8744.6504	1250.243	578496.02	7	8744.5487	B82	11.63
13	5214.7276	1304.6892	498829.11	4	5214.6845	B47	8.25
14	8842.7264	1264.2539	369704.78	7	8843.6171	B83	12.62
15	7111.8373	1423.3747	399446.44	5	7111.7802	Y67	8.03
16	6597.5807	1320.5234	546500.43	5	6597.5051	Y61	11.46
17	9770.3167	1396.7668	275673.87	7	9770.2408	Y92	7.77
18	7040.7991	1409.1671	344665.09	5	7040.7431	Y66	7.96
19	5399.7797	1350.9522	516222.15	4	5399.7646	B49	2.81
20	4720.483	1181.128	332324.21	4	4720.4316	B43	10.88
21	9885.3848	1236.6804	249886.62	8	9885.2677	Y93	11.84
22	8399.4663	1200.931	260604.5	7	8400.3791	B78	10.67
23	4364.2594	1455.7604	331278.2	3	4364.2257	B40	7.74
24	4293.2131	1432.0783	325677.88	3	4293.1885	B39	5.71
25	10720.773	1341.1039	576153.32	8	10720.6906	Y101	7.69

26	9983.4469	1248.9381	752545.25	8	9984.3361	Y94	11.33
27	8329.4621	1190.9304	249450.7	7	8329.342	B77	14.42
28	6952.7637	1391.56	222877.72	5	6953.7111	Y65	7.91
29	10947.8669	1369.4906	292211.36	8	10948.8016	Y103	6.18
30	6114.3102	1223.8693	301314.12	5	6114.2358	Y56	12.17
31	9326.944	1333.4278	154600.59	7	9326.8864	B88	6.17
32	3044.5332	1015.8517	147459.23	3	3044.5131	B28	6.59
33	8956.8241	1280.5536	260667.39	7	8956.7012	B84	13.73
34	3044.5477	1523.2811	145328.79	2	3044.5131	B28	11.37
35	2827.4285	943.4835	139018.91	3	2827.4069	Y25	7.65
36	9457.1457	1183.1505	120045.35	8	9458.0974	Y89	5.36
37	6370.4692	1275.1011	185088.99	5	6369.3941	Y59	11.42
38	6795.6649	1360.1403	295078.35	5	6795.6419	Y63	3.38
39	2759.3677	920.7965	95927.65	3	2759.3443	B25	8.49
40	10832.8488	1355.1134	153498.14	8	10833.7746	Y102	7.06
41	2043.0187	1022.5166	113731.26	2	2042.9949	B19	11.64
42	10447.5178	1306.947	398684.07	8	10448.4693	B99	4.87
43	2258.0983	1130.0564	168146.49	2	2258.0855	B21	5.64
44	9459.192	1352.3204	154345.6	7	9458.0974	Y89	9.75
45	9213.8724	1317.2748	203243.87	7	9213.8023	B87	7.61
46	8201.2747	1367.8864	263486.83	6	8201.247	B76	3.38
47	4607.401	1152.8575	231715.91	4	4607.3476	B42	11.59
48	9884.3469	1413.0568	272974.12	7	9885.2677	Y93	8.25
49	5857.1732	1465.3006	127915.92	4	5857.0983	Y53	12.8
50	4954.6035	1239.6582	142313.56	4	4954.5321	B45	14.42
51	9071.7363	1296.9696	145746.86	7	9071.7281	B85	0.9
52	1159.6001	1160.6074	114108.98	1	1159.5873	B11	11.05

53	2069.0726	1035.5436	66923.15	2	2069.047	Y19	12.4
54	2311.1852	1156.5999	134181.24	2	2311.1736	Y21	4.99
55	2144.0663	1073.0404	85736.58	2	2144.0426	B20	11.05
56	2371.183	791.4016	52272.93	3	2371.1696	B22	5.65
57	1060.5298	1061.5371	76062.54	1	1060.5189	B10	10.28
58	5983.1062	1496.7838	99109.62	4	5983.0248	B55	13.61
59	3191.6008	1064.8742	91704.87	3	3191.5815	B29	6.02
60	1060.6378	531.3262	59989.94	2	1060.6281	Y10	9.12
61	13069.9982	1308.0071	233874.7	10	13069.9526	Y122	3.49
62	9791.229	1224.9109	125600.67	8	9792.12	B93	11.37
63	1767.8482	884.9314	58884.87	2	1767.8315	B17	9.44
64	492.31	493.3173	65126.09	1	492.3061	Y4	8.12
65	6226.3645	1246.2802	147285.67	5	6227.3199	Y57	7.54
66	2945.4674	1473.741	130587.33	2	2945.4447	B27	7.71
67	1159.5954	580.805	48748.33	2	1159.5873	B11	6.98
68	4422.355	1475.1256	83662.06	3	4422.3363	Y40	4.22
69	3897.9609	1949.9877	38682.11	2	3897.9353	B35	6.56
70	3191.6007	1596.8076	81819.49	2	3191.5815	B29	6.01
71	1823.9671	912.9908	38984.18	2	1823.9458	Y16	11.64
72	6114.2737	1529.5757	78755.31	4	6114.2358	Y56	6.19
73	747.47	748.4773	45343.22	1	747.4643	Y7	7.59
74	2182.1466	1092.0806	62289.98	2	2182.1311	Y20	7.13
75	947.4382	948.4454	44991.47	1	947.4349	B9	3.48
76	4026.0783	1343.0334	104126.77	3	4026.0303	B36	11.94
77	4323.2999	1442.1072	112706.28	3	4323.2679	Y39	7.4
78	1696.801	849.4078	43686.78	2	1696.7944	B16	3.87
79	3169.5567	1057.5262	40588.87	3	3169.5608	Y29	-1.3

80	4992.6924	999.5457	30615.38	5	4992.6529	Y45	7.91
81	3354.6599	1678.3372	44408.25	2	3354.6449	B30	4.48
82	1508.8271	755.4208	25726.2	2	1508.8239	Y14	2.12
83	9571.257	1197.4144	41025.31	8	9572.1403	Y90	12.44
84	1982.022	992.0183	35218.33	2	1982.015	Y18	3.55
85	1637.8757	819.9451	14673.77	2	1637.8665	Y15	5.62
86	2846.3756	949.7991	33001.82	3	2846.3763	B26	-0.26
87	3354.6666	1119.2295	59837.12	3	3354.6449	B30	6.47
88	5455.8458	1092.1764	18291.36	5	5456.7861	B50	11.38
89	6484.4232	1622.1131	47019.35	4	6484.421	Y60	0.33
90	3455.7063	1728.8605	20252.84	2	3455.6925	B31	4
91	6696.6078	1117.1086	40657.36	6	6696.5735	Y62	5.12
92	2999.4741	1000.832	14800.31	3	2999.4553	Y27	6.25
93	3098.5422	1033.8547	12598.05	3	3098.5237	Y28	5.96

Table S8. Relative abundance of seven proteins in each batch. The average values were calculated from three technical replicates of each batch, as well as the standard deviation.

Protein ID	Annotation	Abundance in 18-318	Abundance in 19-045	Abundance in 19-056
XP_002493171.1	NAD(+)-dependent formate dehydrogenase	0.018±0.0083	0.005±0.0004	0.017±0.0005
XP_002490864.1	Translational elongation factor EF-1 alpha	0.013±0.0025	0.011±0.0006	0.011±0.0017
XP_002493194.1	Uroporphyrinogen decarboxylase	0.008±0.0018	0.011±0.001	0.011±0.0017
XP_002490697.1	Hexokinase-2	0.013±0.0037	0.005±0.0004	0.008±0.0002
XP_002491701.1	acetate--CoA	0.011±0.0013	0.007±0.0009	0.006±0.0004

	ligase			
XP_002491345.1	Glyceraldehyde-3-phosphate dehydrogenase, isozyme 3	0.006±0.0022	0.003±0.0005	0.006±0.0005
XP_002493065.1	Transketolase	0.006±0.0009	0.009±0.0021	0.008±0.0008

62