Food and Chemical Toxicology 147 (2021) 111888

Contents lists available at ScienceDirect

Food and Chemical Toxicology

journal homepage: www.elsevier.com/locate/foodchemtox

Evaluating potential risks of food allergy of novel food sources based on comparison of proteins predicted from genomes and compared to www.AllergenOnline.org☆

Mohamed Abdelmoteleb [a,b], Chi Zhang [c], Brian Furey [d], Mark Kozubal [d], Hywel Griffiths [e], Marion Champeaud [e], Richard E. Goodman [a,*]

[a] Food Allergy Research and Resource Program, Dept. of Food Science & Technology, University of of Nebraska, 1901 North 21St Street, Lincoln, NE 68588-6207, USA
[b] Department of Botany, Faculty of Science, Mansoura University, Mansoura, 35516, Egypt
[c] School of Biological Sciences, University of Nebraska, Lincoln, NE, 68588, USA
[d] The Fynder Group, Inc. DBA Nature's Fynd, 815 W Pershing Rd Unit 4, Chicago, IL, 60609, USA
[e] Fermentalg, 4 rue Rivière, 33500 Libourne, France

ARTICLE INFO

Keywords:
Chlorella
Galdieria
Fusarium
Bioinformatics
Proteins
Allergy
Cross-reactivity
Genome

ABSTRACT

Potential proteins from three novel food sources (*Chlorella variabilis*, *Galdieria sulphuraria*, and *Fusarium* strain flavolapis) were predicted from genomic sequences and were evaluated for potential risks of allergic cross-reactivity by comparing the predicted amino acid sequences against the allergens in the www.AllergenOnline.org (AOL) database. The preliminary analysis used CODEX Alimentarius limits of >35% identity over 80 amino acids to evaluate the predicted proteins which include many evolutionarily conserved proteins. Regulators might expect clinical serum IgE tests based on identity matches above the criteria if the proteins were introduced in genetically engineered crops. Some regulators have the same expectations for proteins in novel foods. To address the inequality of extensively conserved sequences, we compared the predicted proteins from curated genomes of 23 highly diverse allergenic species from animals, plants and arthropods as well as humans to AOL sequences and compiled identities. Identity matches greater than CODEX limits (>35% ID over 80 AA) are common for many proteins that are conserved through extensive evolution but are not predictive of published allergy risks based on observed taxonomic cross-reactivity. Therefore, we recommend changes in the allergen databases or methods of identifying matches for risk evaluation of new food sources. Our results provide critical data for redefining allergens in AOL or for providing guidance on more predictive sequence identity matches for risk assessment of possible risks of food allergy.

1. Introduction

Allergic reactions to food can pose a serious risk to the health and wellbeing of consumers. Allergen management of commercial packaged foods is through appropriate labeling to warn allergic consumers of the specific contents so they can avoid foods that would put them at risk. The United States, the European Union, and many other countries require labeling of all ingredients, and certainly those viewed as major allergenic sources. Regulations also exist in many countries for managing and identifying potential cross-contact between allergens for foods which do not contain allergens in their ingredient lists. The US recognizes eight major sources as priority allergenic sources (peanut, tree nuts, milk, eggs, crustacean shellfish, finned fish, soybeans and wheat), the European Union recognizes 14 allergenic sources that require labeling in packaged foods with a reduced number of tree nuts, but adding barley, rye and oats to cereals for gluten, mustard, sesame seeds, lupin, molluscan shell fish and the preservative sulfites including sulfur dioxide (Taylor and Hefle, 2006; Muraro et al., 2014).

Abbreviations: AA, amino acid; AOL, AllergenOnline; BLAST, FASTA sequence search tool; NCBI, National Center for Biotechnology Information.

☆ Grant support: This research was funded primarily by the Goodman laboratory at the Food Allergy Research and Resource Program (FARRP) at the University of Nebraska with partial funding from Fermentalg and Nature's Fynd.

* Corresponding author.
E-mail addresses: dmotelb87@gmail.com (M. Abdelmoteleb), zhang.chi@unl.edu (C. Zhang), bfurey@naturesfynd.com (B. Furey), mkozubal@naturesfynd.com (M. Kozubal), hgriffiths@fermentalg.com (H. Griffiths), mchampeaud@fermentalg.com (M. Champeaud), rgoodman2@unl.edu (R.E. Goodman).

https://doi.org/10.1016/j.fct.2020.111888
Received 16 September 2020; Received in revised form 23 November 2020; Accepted 25 November 2020
Available online 1 December 2020

The history of human exposure to the relevant food source and records related to the developed food (novel food source, or gene/protein donor) including the published history of allergy is important in judging potential hazards and risks.

Understanding food allergy risks requires knowledge of the proteins in various foods that commonly or less-commonly cause food allergy as well as the mechanisms of the allergic response. For example, peanut (*Arachis hypogaea*) is a source of severe allergic reactions in many countries. The dominant allergens in peanut are the most abundant seed storage proteins: Ara h 1, vicilin; Ara h 2 and Ara h 6, 2S albumins and Ara h 3, a legumin-like protein (Porterfield et al., 2009; Palladino and Breitender, 2018; Cabanillas et al., 2018). Of these, two are highly soluble proteins that are not rapidly digested at acidic pH by pepsin and so are readily available for immediate reactivity in the mouth or the intestinal tract when consumed. Two are less soluble in water yet are present in sufficient quantities that they still present significant risk. Twelve other peanut proteins are recognized as allergens, though clearly less potent clinically than Ara h 1, Ara h 2, Ara h 3 and Ara h 6. These proteins have been reported to be bound by IgE from some allergic subjects and in some cases when presented at unnatural high abundance in basophil assays, they may stimulate histamine release. The proteins that are low in abundance in the natural food have not been identified as major allergens except possibly the peanut oleosins (Schwager et al., 2017). The majority of proteins from Peanut, however, are not recognized as allergens.

Risks of allergy are dose dependent so identification of a protein as an allergen does not mean it represents a significant risk of food allergy unless it is common and abundant. Risks of allergy also vary markedly among people allergic to the same source (Westerhout et al., 2019). Protein homologues of the dominant peanut allergens are found in other legumes and tree nuts and are the major allergens for most people with clinical allergy to those sources (Cabanillas et al., 2018). Proteins that cause cross-reactions can usually be grouped into protein families, although there are many non-allergic proteins within any of the identified biochemical protein groups. For example, the important muscle allergen tropomyosin from crustaceans is highly conserved. The sequence homology between allergenic crustaceans, mollusks and insects such as mealworm is over 60% identity by BLASTP or FASTA and there is IgE cross-reactivity from the proteins of these organisms using sera from many shrimp allergic subjects. However, homologues in birds and mammals including humans are more than 52% identical to shrimp tropomyosin and while some *in vitro* IgE cross-reactivity is observed for some subjects' sera, there is little evidence of shared allergy (Faber et al., 2017; Ruethers et al., 2018).

Novel food ingredient sources are being developed to meet the growing demand for dietary proteins in industrialized countries due to the increasing human population, concerns for animal welfare, and environmental impacts of traditional sources of protein (Bleakley and Hayes, 2017; Frigerio et al., 2020). Many diverse food sources have been consumed in some geographic regions with a history of safe use, although the use and safety or risk are rarely well documented in less industrially developed regions. Some potential food sources are truly novel, with no history of safe human consumption including microbial sources such as specific microalgae, fungi or yeasts as whole foods or ingredients. Since there are no validated methods for predicting *de novo* sensitization, the allergenicity assessment for these truly novel foods is focused on immediate risks to consumers due to the presence of existing IgE that could arise either from unexpected exposure to an allergen to which they are already allergic, or to a likely cross-reactive protein. A sound risk assessment process will have the primary focus on judging knowledge of history of allergy to the source, and similarity of the proteins of the source to known allergens.

The safety assessment of genetically engineered (GE) organisms has served as a model for assessing allergenicity risk of some new foods in the United States (US). Hazard identification and risk assessment steps for GE organisms were broadly discussed in the early 1990's (Federal Register Docket No. 92N-0139, Vol 57, No. 104, May 29, 1992) and (Metcalfe et al., 1996). A primary health related concern has always been whether a new gene in a GE organism encodes an allergen or a potentially cross-reactive protein that would act as an allergen for those who are already allergic. Advisory groups were convened by the Food an Agricultural Organization (FAO) and World Health Organization (WHO) panels in 1996 and 2000. In 2001 the FAO/WHO held a meeting and recommended untested steps including looking for peptide matches of 6 contiguous amino acids and targeted serum IgE binding studies (FAO/WHO, 2001). In 1996 only a few hundred allergenic protein amino acid (AA) sequences were known in publications. The AA sequences of the new protein in the GE crops were compared to allergens in private databases of the developer, or in the NCBI Protein non-redundant (nr) database using keyword search limits. Searches were accomplished by FASTA in small databases or BLASTP in NCBI (Pearson, WR, 2000; Pearson WR, 2014). Alignments that might represent an allergen were searched for identity matches of eight contiguous amino acids to any segment of any allergen. If matched, serum IgE binding tests would be conducted focusing on those with allergies to the source of the new protein. However, in practical terms, developers often abandon those as potential products.

Evaluation of the short segment amino acid comparisons (6–8 amino acid matches) were later shown to be non-predictive (Hileman et al., 2002). The CODEX Alimentarius meeting in 2001 as published in 2003 and reaffirmed in 2009 (CODEX, CAC/GL 44 in 2003 and reviewed in 2009 (CODEX Alimentarius Commission, 2009)) considered those criteria and other information and the consensus was that a FASTA search looking for minimum identity matches of >35% over 80 amino acids was a more predictive test (Goodman et al., 2008).

It has been suggested that the current CODEX guideline of >35% identity over at least 80 amino acids threshold be considered in conjunction with *E*-scores (expectation scores) generated from the FASTA algorithm to make a more informed decision as to whether a protein has the potential to cause allergenic cross-reactivity (Thomas et al., 2005; Ladics et al., 2007; Silvanovich et al., 2009; Cressman et al., 2009). The *E*-score reflects the measure of relatedness among protein sequences and can help separate the potential random occurrence of aligned sequences from those alignments that may share structurally relevant similarities. A small *E*-score (e.g., less than 1e-7) reflects a likely functional similarity and may suggest a biologically relevant similarity for allergy or potential cross-reactivity, while large *E*-scores (>1.0) are typically associated with alignments that do not represent a biologically relevant similarity (Pearson 2000, 2014, 2016; Henikoff and Henikoff 1992, 1996).

However, this guidance should be viewed as highly conservative and precautionary based on historical experiences of cross-reactivity and clinical co-reactivity. Clinically important IgE cross reactivity is common for proteins sharing >70% AA identity over nearly their full-lengths, yet cross-reactivity is extremely rare for proteins sharing less than 50% identity (Aalberse, 2000). Other aspects of protein structure and IgE binding are important to consider cross-reactivity (Aalberse et al., 2001).

The AllergenOnline.org (AOL) database at the Food Allergy Research and Resource Program (FARRP) at the University of Nebraska was started in 2004–2005. It is a public, peer-reviewed database of allergens based on protein AA sequences in the NCBI Protein database following evaluation of published evidence in peer-reviewed literature (Goodman et al., 2005, 2016). The AOL database includes proteins from studies of airway, contact, food, venom and salivary allergen sources with IgE binding. When provided in publications, evidence of histamine release and clinical reactivity adds confidence to calling the proteins an allergen. The AllergenOnline.org database has been updated annually by adding newly published allergens every year from 2006 through 2020 by a review process with a panel of allergen experts that include researchers and clinicians (Goodman et al., 2016). It has been used for evaluating risks of food allergy for many GE crops and can be used for evaluating new foods.

AOL uses FASTA comparison with the criteria of matches being >35% identity over 80 amino acids as was set by the CODEX Allergenicity guideline in 2003. But since some proteins or alignments might be of less than 80AA if fragments of allergens were transferred into other species, or if these sections contain high identity segments that could cause severe cross-reactivity AOL also adjusts the calculation with normalization of alignments less than 80 AA. As an example an N-terminal segment of 77AA of Ara h 2 includes two or three IgE binding epitopes and if transferred to a non-peanut food could cause severe clinical reactions in some peanut allergic consumers (Dreskin et al., 2019). As described online (www.allergenonline.org) in a support page for sequence searches, the number of AA identity matches of any alignment less than 80 AA is recalculated by dividing by 0.80 to normalize to an 80 AA length. The minimum identity match to consider as possibly cross-reactivity is 29 identical AA in any FASTA alignment which is calculated as 36.25%. This modified FASTA search provides a more reliable evaluation of potential risks than either a strict FASTA search eliminating sequences shorter than 80 AA or a short (8 AA) alignment.

For truly new foods it is now possible to use modern techniques of proteomics and genomics to predict all potential proteins from new sources. Evaluating all proven proteins of a whole organism for potential risks of food allergy would not be efficient or effective if that required identification of each individual protein in the food with tests of possible IgE binding, or clinical reactivity. Therefore, evaluation of potential risks of food allergy from an organism such as an alga, fungus or new plant that does not have a history of human consumption requires new evaluation steps. Some regulators and scientific advisors have recommended using predicted proteins from the whole organism's genome or transcriptome for comparison to allergen databases using the CODEX guidelines to predict risks of food allergy. Importantly, the CODEX guideline was not intended to evaluate the full-proteome or predicted protein dataset of a whole organism as the criteria of >35% identity over 80 has not been validated for whole proteome comparisons.

The end-result of the bioinformatics comparison of proteins with allergens is a decision about the need for specific serum testing and if so, the specific allergic population that should be used to collect serum samples (Goodman et al., 2005). But, since appropriate serum testing is not trivial, correct interpretation of bioinformatics findings are important. Many genes and their expressed proteins, including many genes that encode "minor" allergens are highly conserved across species and so it is highly probable that these will trigger a match using the CODEX guidelines. Predictions of protein sequences from genomic and transcriptomic evaluations therefore require quality checks to understand relevance before deciding on the need for clinical testing and critical evaluation of the criteria used for decision making is required (Siruguri et al., 2015).

Based on our years of use and development of AllergenOnline.org, it appears that the CODEX guidelines are far too conservative to judge proteins that match evolutionarily conserved allergens, especially when applied to whole genomes. We have therefore performed this study in part to understand the extent of over-predictions. We have evaluated protein sequence identity matches between three diverse species (a green alga *Chlorella sp.*, a red alga *Galdieria sulpharuraia* and a *Fusarium* strain flavolapis) searching the AllergenOnline.org (AOL) database and the NCBI Protein database to consider matches to likely allergens.

1.1. There are three objectives in this study

First, to evaluate identities of all possible proteins from the genomes of three species intended for food use based on comparison of the predicted proteins against allergens in the AllergenOnline.org database using the CODEX guidelines of >35% identity over 80 AA.

Second, to address the inequality of extensively conserved sequences, we compared the predicted proteins from the genomes of 23 highly diverse allergenic and non-allergenic species; including human, animals, plants and arthropods to all AOL sequences, compiled identities to understand how common high identity matches are, and evaluated patterns of identity across protein types.

Third, to critically evaluate the limits of the CODEX guidelines when used as a whole genome analysis, using all types of proteins. The overall goal being to determine what in addition to the CODEX criteria is reasonable for risk assessment of whole foods.

1.2. Tests of three species based on genomic predictions of proteins

We chose to use a green alga *Chlorella variabilis,* a red alga *Galdieria sulphuraria*, and a newly identified *Fusarium* strain flavolapis fungus as test organisms. These organisms are being developed as single-cell food protein resources. *Chlorella* is a genus of single-celled green algae which contains high concentrations of protein (51%–60% of dry matter), amino acids, vitamins, dietary fiber, and a variety of antioxidants, bioactive materials, and chlorophylls. Green algae have a history of sustainable production and consumption. (Klamcyzynska and Mooney, 2017). *Chlorella vulgaris* and *Chlorella pyrenoidosa* are not considered novel in the EU since they have been historically consumed by humans (Regulation EC No. 258/97). In the US they are recognized as GRAS by the FDA as algae commonly consumed in foods in many countries (Wells et al., 2017). Recently the genome of *Chlorella variabilis,* NC64A was completed and was used here as a model genome (Blanc et al., 2010).

The unicellular red algae, *Galdieria sulphuraria,* was isolated by developers from extreme environments (from pH 0 to 4, and up to 56 °C) and has been proposed as an edible alga with a high content of protein and other dietary important nutrients. This alga can be grown via fermentation and is being developed for use in food products (Schonknecht et al., 2013), but has not yet been consumed by humans.

A single species of *Fusarium* is already used in several food products with the brand name, Quorn. Quorn is produced and marketed as a human food by Marlow Foods, Ltd. Quorn foods contain mycoprotein which is derived from *Fusarium venenatum*, which is grown by fermentation (Finnigan et al., 2019). Products of Quorn have been consumed as a non-meat protein source in the United Kingdom for 30 years and since 2002 in the US. There are a few case reports of food allergy to Quorn (Katona and Kaminski, 2002; Hoff et al., 2003a). Some of those may be due to inhalation allergy to proteins of *Fusarium sp.* (Weber and Levetin, 2014). Some consumers of Quorn have experienced transient GI symptoms without IgE antibody production. A very small number have experienced possible IgE mediated food allergic reactions including one reported fatal reaction (Tee et al., 1993; Hoff et al., 2003a, 2003b; Yeh et al., 2016; Jacobson and DePorter, 2018). To put this in perspective, many common food sources have caused at least one fatal food allergic reaction and as long as packaged food is labeled clearly, consumers with allergies can avoid consumption of foods that may cause allergic reactions if they are properly labeled (Ramsey et al., 2019; Gowland and Walker, 2015). Other strains or species of *Fusarium* with different compositions are now under development as possible food sources including, *Fusarium* strain flavolapis, the strain we are using here for which the developers have performed whole genome sequencing.

2. Methods

2.1. Preparation of protein sequences of the three targeted genomes

The predicted proteins for the genome of *Chlorella variabilis* NC64A were downloaded from the NCBI genome library (https://www.ncbi.nlm.nih.gov/genome/?term=Chlorella+variabilis+ %5Borgn%5D). For *Galdieria sulphuraria*, the company Fermentalg provided the DNA sequences which were identified using Illumina sequencing (2x150 bp reads). The sequencing quality was checked using FastQC (Andrews 2010) and cleaned using PRINSEQ (prinseq.sourceforge.net) by trimming off low quality bases. Two assemblers were used, SPAdes with 21, 33, 55 and 77 k-mer values (Bankevich et al., 2012), and Trinity using 25

k-mer (https://github.com/trinityrnaseq/trinityrnaseq/wiki). Post assembly polishing was performed using Pilon (Walker et al., 2014). The quality of assembly was checked using Quast (Gurevich et al., 2013). The percentage of mapping was evaluated using BWA mapper (Li and Durbin, 2009). Genes were predicted using the *Galdieria* model from AUGUSTUS (Stanke and Morgenstern, 2005). Sequences to exclude included tRNA sequences which were predicted using tRNAscan-SE (Lowe and Eddy, 1997) and rRNA which were predicted using barrnap (https://github.com/tseemann/barrnap). Functional annotation was conducted by a combination of AUGUSTUS software and BLASTP comparison for the predicted proteins against the published *Galdieria sulphuraria* genome (https://www.ncbi.nlm.nih.-gov/genome/?term=Galdieria+sulphuraria) from the NCBI library. Sequences were compiled into FASTA format files for comparison to the AllergenOnline.org database. The compiled sequences were also compared to the published *Galdieria sulphuraria* genomic sequences filed by Schonknecht et al., as described in 2013 as ASM34128v1 using alignment tools in order to check for potential sources of inaccuracy.

Nature's Fynd provided the genomic sequences for *Fusarium* strain flavolapis, which they are developing for use as a food ingredient. They performed genomic sequencing using Pacbio (for long-reads) and Illumina (2x250 bp reads) for short, high quality reads of this cultured species. These sequences were compiled and evaluated for accuracy and completeness using FASTQC. Sequences were compiled using assemblers MaSuRCA with 22 k-mer value (Zimin et al., 2013) and SPAdes (Bankevich et al., 2012) used K-mers of 21, 33, 55, 77, 99 and 127. Post assembly polishing used Pilon (Walker et al., 2014). Pacbio reads were mapped using Minimap2 (Li 2016), and Illumina reads were mapped using Bowtie2 (Langmead and Salzberg, 2012). Genes were predicted using the *Fusarium* model from AUGUSTUS (King et al., 2015), mitochondrial genes were predicted using Prodigal (Hyatt et al., 2010), tRNA were predicted using tRNA scan-SE (Lowe and Chan, 2016) and rRNA were predicted using Barrnap software (https://github.com/tseeman/barrnap/). Functional annotation was done using the ERGO software package of IgenBio (Wilder et al., 2016). The overall sequence completeness was further evaluated by comparison to the genomes of strains of *Fusarium* sp. which had been previously characterized to provide a framework for understanding completeness (Niehaus et al., 2016).

To provide reasonable comparisons, the predicted proteins for the genomes of 23 species representing foods of diverse allergenic risks and included those of human, other animals and plants. The sequences were downloaded from public databases including the NCBI genome library (https://www.ncbi.nlm.nih.gov/genome), EnsemblPlants (http://plants.ensembl.org/index .html), and Phytozome V. 12, the Plant Genomics Resource (https://phytozome.jgi.doe.gov/pz/portal.html#) as summarized in Table 1. For species without published genomes as of October 2018, we downloaded the predicted protein sequences from the NCBI protein library. All protein sequences were downloaded on October 2018. The bioinformatics pipeline was completed using our lab cluster on the Holland Computer Center server at the University of Nebraska.

Table 1
Sources for predicted protein sequences from 23 genomes of diverse species.

Species	Source
Human (*Homo sapiens*)	ftp://ftp.ncbi.nih.gov/genomes/H sapiens/protein/
Baker's yeast (*Saccharomyces cerevisiae*)	http://downloads.yeastgenome.org/sequence/S288C.reference/orfprotein/
Candida albicans SC5314	http://www.candidagenome.org/download/sequence/C.albicans.SC5314/Assembly22/current/
Cod (*Gadus morhua*)	ftp://ftp.ensembl.org/pub/release-86/fasta/gadus.morhua/pep/
Chicken (*Gallus gallus*)	ftp://ftp.ensembl.org/pub/release-86/fasta/gallus.gallus/pep/
Bovine (*Bos taurus*)	ftp://ftp.ensembl.org/pub/release-86/fasta/bos.taurus/pep/
Drosophila melanogaster	ftp://ftp.flybase.net/genomes/Drosophila.melanogaster/dmel.r6.09.FB2016.01/fasta/
Salmon (*Salmo salar*)	ftp://ftp.ncbi.nih.gov/genomes/Salmo.salar/protein/
Papaya (*Carica papaya*)	ftp://ftp.ncbi.nih.gov/genomes/Carica.papaya/protein/
Soybeans (*Glycine max*)	ftp://ftp.ncbi.nih.gov/genomes/Glycine.max/protein/
Apple (*Malus domestica*)	ftp://ftp.ncbi.nih.gov/genomes/Malus.domestica/protein/
Rice (*Oryza sativa*)	ftp://ftp.ncbi.nih.gov/genomes/Oryza.sativa.Japonica.Group/protein/
Peanut (*Arachis hypogaea*)	ftp://ftp.ncbi.nih.gov/genomes/Arachis.hypogaea/protein/
Peach (*Prunus persica*)	ftp://ftp.ncbi.nih.gov/genomes/Prunus persica/protein/
Beans (*Phaseolus vulgaris*)	https://www.ncbi.nlm.nih.gov/genome/?term=Phaseolus+vulgaris+%5Borgn%5D
Potato (*Solanum tuberosum*)	ftp://ftp.ncbi.nih.gov/genomes/Solanum.tuberosum/protein/
Wheat (*Triticum aestivum*)	https://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org.Taestivum.er
Maize (*Zea mays*)	ftp://ftp.ncbi.nih.gov/genomes/Zea.mays/protein/
Arabidopsis thaliana	https://www.ncbi.nlm.nih.gov/genome/?term=arabidopsis++thaliana+%5Borgn%5D
Almond (*Prunus dulcis*)[a]	https://www.ncbi.nlm.nih.gov/protein/?term=prunus+dulcis
Pecan (*Carya illinoinensis*)[a]	https://www.ncbi.nlm.nih.gov/protein/?term=carya+illinoinensis
Pistachio (*Pistacia vera*)[a]	https://www.ncbi.nlm.nih.gov/protein/?term=pistacia+vera+%5Borgn%5D
English Walnut (*Juglans regia*)	ftp://ftp.ncbi.nih.gov/genomes/Juglans.regia/protein/

[a] Species without complete published genomes before October 2018.

2.2. FASTA comparison for the predicted protein sequences of the genomes were compared to Allergenonline.org version 16 and 18B

Predicted protein sequences from the proposed three novel food species and 23 diverse species were compared to allergens in versions 16 and 18B of www.AllergenOnline.org by overall FASTA 35. FASTA version 35 was installed on the Holland Computing Center server to allow batch searches that mimic the individual protein searches available on our AllergenOnline.org website, however based on the best identity matches over 80 AA long. Different *E*-score thresholds (10, 1, 0.001, 1e-7, 1e-30, 1e-50, 1e-75, 1e-100) were used to check the significance of matches on the private HCC searches. The same scoring matrix was used (BLOSUM 50) as on the public AllergenOnline.org database. The sequence matches to proteins in AllergenOnline.org were compiled in an Excel worksheet with a record of the highest match identity. The resulted matches were evaluated to identify matches of >35% identity over 80 or more amino acid segments.

2.3. BLASTP comparison of predicted protein sequences within the NCBI non-redundant protein sequences database that includes annotated protein sequences from GenBank, RefSeq and TPA as well as SwissProt, PIR, PRF and PDB

Predicted protein sequences of *Chlorella variabilis*, *Galdieria* sp. and *Fusarium* strain flavolapis. as well as the 23 other species used in this study were used to search the general protein database using the current version of BLASTP in 2018 and early in 2019. The website is https://blast.ncbi.nlm.nih.gov/BLAST.cgi. The current version of BLASTP outputs changed markedly in July 2019, removing the ability to use keyword limits in BLASTP searches to restrict matches to particular categories of sequences based on keywords. In addition, the output of BLASTP has changed and we used the Traditional Results for historical comparisons. Searches without keyword limits allows the highest identity matches to be viewed for evaluation of the common conservation of the protein sequences. The previous selection criteria using keyword limits such as "allergy" or "allergen" were removed. Those changes speed the searches but eliminates useful screening decisions. We also used BLASTP searches

of species targets from the 23 species and of the matched allergens from out AllergenOnline.org to provide guidance on the relevance of low-identity matches including >35% identity over 80 amino acids.

3. Results

3.1. Prediction of Galdieria sulphuraria. And Fusarium strain flavolapis proteins based on genomic DNA sequences

For *Galdieria sulphuraria*, the number of reads after checking quality and trimming are 26.4 Mbp. Assembly metrics are: 1998 contigs, largest contig 294001B, N50 54420B, N75 16958B, L50 66, L75 164, and GC% 40.27 for SPAdes; and for Trinity are 2890 contigs, largest contig 154130B, N50 24717B, N75 11797B, L50 292, L75 677, and GC% 40.30. The reads were mapped at 99.67% for SPAdes, and 99.59% for Trinity. The number of predicted proteins for *Galdieria sulphuraria* was 5701 from SPAdes and 11976 from Trinity.

For *Fusarium* strain flavolapis, the quality of the sequences included 340k reads after trimming and correcting from Pacbio, and 56.5 M read pairs from Illumina. Assembled sequences included 89 contigs, with the largest contig being 4.9 MB, N50 for 3.2 MB, N75 for 2.3 MB and L50 6, L75 10 and 0 Ns with a GC content of 48.3%. Pacbio reads mapped at 99.95% using Minimap2 software. Illumina reads mapped at 99.81% using Bowtie2 software. The number of predicted proteins were 14239.

3.2. Comparison of all possible proteins from the genome of the three novel foods against allergens in AOL

The total number of unique matches to allergens for predicted proteins from the three potential food species that scored over a range of *E*-scores with results >35% identity limit over 80 AA of CODEX guidelines are shown in Table 2. The normal default *E* score for FASTA or for BLAST is 10, but smaller *E* score numbers restrict the output to provide more stringent alignments. The purpose of these comparisons was to evaluate whether the CODEX criteria are reasonable for risk assessment of the three proteins using >35% identity over 80 AA as the criteria to bench mark a need for serum IgE tests or other additional evaluations. As shown in Table 2, the three species of interest have not been consumed (widely) by humans and are thus not known to cause allergies, yet they show very high numbers of matches greater than 35% identity over 80 AA at 1e-07 to allergens in AOL, with *E* score settings much smaller than the default of BLASTP. More realistic numbers of alignments, meaning identities between species that have been reported as possibly being cross-reactive were found when the *E*-score was set to 1e-100.

For comparison, we tested all predicted proteins from the genomes of 23 species ranging from humans to fungi, fish, mammals and many species of plants to evaluate the number of possible risky proteins. These matches are summarized in Table 3 for comparison to the three species of interest. Matches following CODEX guidelines are intended to identify proteins that may be sufficiently similar to an allergen to suspect possible IgE cross-reactivity and the possibility of triggering a clinically important allergic reaction. As shown in Table 3, a significant number of matches >35% identity to multiple allergens was found for proteins from all 23 species with *E* scores of 10 or even 1. Even using an *E*-score of 1e-100, the number of any match unique proteins seems far higher than expected based on numbers of allergenic proteins in commonly allergenic sources. Experiences in clinical research demonstrates that even the most commonly allergenic species such as peanut, produce only 4 to 6 commonly allergenic unique proteins (Ara h 1, Ara h 2, Ara h 3, Ara h 6 and possibly Ara h 8 and Ara h 9) and a total of <20 total allergens. Many other commonly allergenic species, such as shrimp list fewer than 10 allergenic proteins that elicit symptoms from human exposure by the airway, contact or ingestion allergens (www.allergen.org). A few sources of airway allergy such as the common house dust mite (HDM) *Dermatophagoides farina* and the evolutionarily related *Dermatophagoides pteronyssinus* have nearly 40 different proteins that may be bound by IgE of people with inhalation allergies. However only three proteins from HDM (Der f 1, Der f 2 and Der f 23, or Der p 1, Der p 2 and Der p 23) are considered major allergens and four others (Der f 4, Der f 5, Der f 7 and Der f 21) are considered mid-level allergens (Thomas, 2015). The other HDM proteins are unlikely to be clinically important because of low level expression, high instability, and unlikely inhalation exposure. Interestingly many of the allergens that have been identified are commonly conserved proteins that share high identity scores across relatively unrelated taxa such as profilins, heat shock proteins and beta-expansins. There are rare to fairly common reports of allergy to some of these species, while only a few clear reports of allergy are common for many species. Our intent in testing 23 species including human proteins was to identify an *E*-score limit that might be valuable for risk assessment and also to test percent identity scores that might be more predictive than the CODEX limit of >35% identity over 80 AA and to consider the relevance of >35% identity.

3.3. Identities of all possible proteins from the genome of the three novel food sources and 23 common species matches to AOL

The results in Table 2 illustrate that the algae (*Chlorella variabilis* NC64A) has sequence matches of >35% identity to between 14 and 991 unique proteins in AOL, depending on which *E*-score limit was used. Even at the moderate *E*-score of 1e-7 there were 159 proteins that suggest potential cross-reactivity. By comparing all predicted proteins from the 23 diverse species including humans (*Homo sapiens*) in Table 3, we found similarly high numbers of matches of the predicted proteins to allergens across the species. Pistachio had the lowest number of matches, but few total proteins have been predicted from nucleotide sequences for pistachio or pecan (Table 3). When we compared the highest scoring aligned proteins of *Chlorella variabilis* to all proteins in AOL version 18B as shown in Table 4. The highest scoring matched allergen was to cyclophilin of *Daucus carota,* but that protein is highly conserved to sequences in all 23 species. Heat shock protein 70 of the *Aedes aegypti* mosquito is highly conserved as shown by sequence matches to proteins in 22 species. The lowest scoring matches in Table 4 include a few bona fide allergens with identity matches close to 35% identity, and with modest *E*-scores. Those include matches to thioredoxin of fungi at 39–40% identity and venom allergen 5 of a wasp at 35.8% identity. Most of the matched allergens are conserved across many species of the 23 chosen here. Many are house-keeping proteins including cyclophilins, heat shock proteins, 60S ribosomal protein, triosephosphate isomerase, aldolase, gliadins. However, the percent identities are not high compared to BLASTP matches to homologues

Table 2
Total number of matches and unique matches to allergens in AOL (>35% sequence identity over 80 AA alignment length) at different *E* scores in the three novel food sources.

Species	Subject Hits	10	1	0.001	1e-7	1e-30	1e-50	1e-75	1e-100
Chlorella variabilis NC64A	Total	277988	82613	9043	3201	413	119	57	35
	Unique	991	752	297	159	64	39	21	14
Galdieria sulphuraria	Total	67989	17792	3202	1222	170	97	50	32
	Unique	101	96	85	73	39	32	12	8
Fusarium strain flavolapis	Total	192772	65321	13320	5867	646	317	135	88
	Unique	508	466	326	232	125	95	44	30

Table 3
Total and unique matches to allergens in AOL for predicted proteins from 23 different allergenic and non-allergenic species.

Species	Subject Hits	10	1	0.001	1e-7	1e-30	1e-50	1e-75	1e-100
Homo sapiens	Total	6200050	2460980	510958	175239	19346	7860	2516	1817
(Human)	Unique	14997	13534	8546	5565	2556	1538	912	557
Saccharomyces cerevisiae	Total	71691	24440	5320	2043	384	243	200	158
(Baker's yeast)	Unique	225	214	164	132	68	52	40	32
Candida albicans SC5314	Total	185065	73070	18846	7712	599	292	174	140
(Yeast)	Unique	648	621	482	327	113	75	45	39
Gadus morhua	Total	339873	118495	24766	10932	1910	991	354	248
(Cod)	Unique	850	806	638	502	268	182	108	72
Bos Taurus	Total	431730	162370	33305	13860	2131	760	350	245
(Bovine)	Unique	1280	1190	865	680	356	227	125	71
Gallus gallus	Total	1067198	463688	112614	41907	6624	3397	865	450
(Chicken)	Unique	2964	2731	1798	1261	636	423	269	153
Drosophila melanogaster	Total	735514	325747	85437	35174	3969	2413	1037	566
(Fruit fly)	Unique	3180	2959	2045	1306	503	286	168	117
Salmo salar	Total	2105661	931620	240600	93818	11910	5695	1318	973
(Salmon)	Unique	7039	6489	4416	2892	1217	720	487	320
Carica papaya	Total	330257	113307	30307	16765	5066	2665	621	149
(Papaya)	Unique	1140	1097	991	877	501	363	175	69
Glycine max	Total	916939	324720	85635	46849	12620	6459	1760	523
(Soybeans)	Unique	3055	2951	2612	2208	1250	881	407	179
Malus domestica	Total	745067	263553	74541	41863	13796	5996	1614	484
(Apple)	Unique	2867	2760	2432	2037	1039	720	320	146
Oryza sativa	Total	612090	174766	30203	17632	5038	2488	648	279
(Rice)	Unique	1710	1578	1255	981	523	328	163	63
Arachis hypogaea	Total	1193850	414633	109245	59692	15021	8476	2356	739
(Peanut)	Unique	4175	4033	3529	2971	1506	1076	486	218
Prunus persica	Total	422277	157454	45557	26298	10252	5115	1433	346
(Peach)	Unique	1701	1637	1416	1201	713	517	264	111
Phaseolus vulgaris	Total	451134	149740	42236	25113	7048	3626	1005	265
(Beans)	Unique	1548	1485	1346	1181	701	488	220	89
Solanum tuberosum	Total	462504	171829	50277	27881	7858	4100	1000	294
(Potato)	Unique	1880	1822	1626	1374	723	512	242	86
Triticum aestivum	Total	5068723	1295317	213159	112949	25380	9739	2927	1436
(Wheat)	Unique	9064	8557	7331	6267	3290	1904	799	384
Zea mays	Total	1126007	346921	60378	30418	9059	4833	1156	528
(Maize)	Unique	3094	2869	2208	1661	813	574	242	127
Arabidopsis thaliana	Total	692802	240908	61433	30158	8702	4575	1083	292
(Mustard)	Unique	2293	2205	1911	1618	834	613	283	112
Prunus dulcis	Total	13102	4540	2619	2323	699	392	26	4
(Almond)	Unique	54	54	52	50	45	25	15	5
Carya illinoinensis	Total	5086	2303	1273	796	440	301	74	52
(Pecan)	Unique	32	32	32	20	17	15	13	13
Pistacia vera	Total	3755	729	285	245	126	42	21	11
(Pistachio)	Unique	8	8	8	8	7	7	7	6
Juglans regia	Total	666338	235167	66984	36964	11699	6744	1573	386
(English Walnut)	Unique	2592	2505	2291	1933	1006	723	343	138

from a variety of protein sources and from species that are not likely to represent risks. For example, BLASTP comparison of triosephosphate isomerase (EFN53775.1) in *Chlorella variabilis* to non-redundant protein database had the top 100 matches to triosephosphate isomerase in diverse species with sequence identity ranged from 69 to 100%. Similarly, *Chlorella* heat shock protein 70 (EFN57963.1) had matches to heat chock proteins in different species with sequence identities of 77.5–100%. This shows the conservation of these proteins among diverse allergenic and non-allergenic species.

Similarly, Table 5 shows that *Galdieria sulphuraria* had matches to 59 weak or putative allergens and 6 very weak matches to food allergens (tropomyosin, vicilin, and convicilin) with *E*-scores >0.02. Due to high sequence identity of evolutionary homologues, these identity matches were over predictive for possible risks of allergic cross-reactivity. The searches were rerun using an *E* score of 1e-7 that removed proteins that are clearly unlikely to cause cross reactive. The results are shown in Table 5. The identified food allergens represent important protein classes of allergens, yet the identity matches shown in this study show very low identities of proteins as with those from *Chlorella*, meaning they are unlikely to be significant risks for cross-reactivity. That can be demonstrated by comparing the matched allergens to the NCBI Protein database using BLASTP. The results for FASTA comparison of predicted proteins of the Quorn fungal genome-predicted proteome, another species of *Fusarium,* was tested for background evaluation. Quorn has been used as a food source in the United Kingdom for >30 years. The results are shown in Supplementary Table 1, that identified 181 matches to weak or putative allergens and 12 very low identity matches to food allergens with very low sequence identity over short AA segments.

3.4. summary examples of FASTA comparisons using all predicted proteins from the 23 studied species

Predicted proteins from the public genomes of all 23 species were compared to AllergenOnline.org looking for matches of >35% identity, using an *E* score cutoff of 1e-07. Wheat genome predicted proteins matched 312 putative allergens, but only eight major allergens. Soybean genome predicted proteins matched 243 putative allergens and 32 matches to major allergens (vicilins and conglycinins of soybean, walnut, pecan and pistachio). Genome predicted human proteins matched 206 weak or putative allergens, one matched the major allergen tropomyosins from a variety of sources including crustacean allergens and those of fruit flies (Drosophila sp.), fish (salmon and cod). Another human protein matched lipid transfer proteins (LTP) with a modest identity match to LTP from pomegranate (42.3% identity with an *E* score of 3.7e-19). Searching AllergenOnline.org with the pomegranate LTP shows many higher identity matches, often >55% ID with *E* scores of

Table 4
FASTA comparison of predicted proteins of *Chlorella variabilis NC64A* compared to AOL V18B (*E*-score: 1e-07). The amino acid sequences of all proteins predicted from the genome of this species were used to search version 18B of the AllergenOnline.org database to find identity matches with proteins listed as allergens or putative allergens in the database using full-length FASTA searches to identify matches of >35% identity with different *E* scores, those from matches at 1e-7 are shown here.

AllergenOnline Version 18B	Highest %Seq_id	Align length	E-score	# of 23 species with matches >35% ID over 80AA
gid|1941|cyclophilin [Daucus carota]	78.8	170	9.00e-75	20
gid|1926|cyclophilin [Catharanthus roseus]	76.8	168	2.70e-54	18
gid|2708|heat shock cognate 70 [*Aedes aegypti*]	73.4	305	4.70e-103	22
gid|2591|heat shock-like protein [Tyrophagus putrescentiae]	73.3	659	6.10e-168	22
gid|2291|Der f 33 allergen [Dermatophagoides farinae]	73.2	455	4.20e-155	23
gid|166|triosephosphat-isomerase [*Triticum aestivum*]	72.6	248	2.10e-105	14
gid|2301|glyceraldehyde-3-phosphate dehydrogenase [*Triticum aestivum*]	70.7	334	1.40e-100	21
gid|338|60S ribosomal protein L3 (Allergen Asp f 23) [Aspergillus fumigatus]	67.1	386	2.00e-118	22
gid|1033|cytochrome_c_ [Curvularia_lunata]	66	103	1.5e-30	
gid|863|cyclophilin [Aspergillus fumigatus]	64.6	161	4.10e-47	18
gid|706|Lactoylglutathione lyase (Methylglyoxalase) (Aldoketomutase) (Glyoxalase I) (Glx I) (Ketone-aldehyde mutase) (S-D-lactoylglutathione methylglyoxal lyase) (Allergen Ory s ?) (Allergen Glb33) (PP33) [Oryza sativa]	62.9	283	1.00e-40	13
gid|543|60S acidic ribosomal protein P2 [Fusarium culmorum]	62.4	109	2.50e-23	14
gid|2076|heat shock protein 70 [Dermatophagoides farinae]	59.6	401	1.80e-71	10
gid|1092|manganese superoxide dismutase-like protein [Pistacia vera]	58.4	202	4.60e-54	17
gid|848|60S acidic ribosomal P1 phosphoprotein Pen b 26 [Penicillium brevicompactum]	57.6	85	1.20e-11	6
gid|648|major allergenic protein Mal f4 [Malassezia furfur]	57.5	320	2.60e-89	20
gid|2255|putative chitinase [Musa acuminata]	56.7	261	1.50e-65	13
gid|1707|aldolase A [*Thunnus albacares*]	56.4	353	7.70e-77	19
gid|587|Chain A, Latex Profilin Hevb8 [Hevea brasiliensis]	56.1	132	2.80e-35	1
gid|489|putative nuclear transport factor 2 [Davidiella tassiana]	55.4	112	5.90e-25	14
gid|2592|aldehyde dehydrogenase-like protein [Tyrophagus putrescentiae]	54.8	489	1.50e-89	20
	54.3	129	1.30e-43	21
gid|1248|eukaryotic translation initiation factor [Forcipomyia taiwana]				
gid|2463|ElF1 superfamily transcriptions factor [*Triticum aestivum*]	54.3	81	1.90e-22	19
gid|2262|transaldolase [Penicillium chrysogenum]	51.4	313	2.70e-74	3
gid|1960|aldolase a, fructose-bisphosphate 1 [*Salmo salar*]	50.6	350	9.50e-68	18
gid|509|group 15 allergen protein [Dermatophagoides farinae]	50	120	9.30e-12	21
gid|651|allergen [Malassezia sympodialis]	50	140	2.60e-27	20
gid|64|Minor allergen Alt a 7 (Alt a VII) [*Alternaria alternata*]	50	200	3.20e-42	15
gid|126|minor allergen beta-fructofuranosidase precursor [Lycopersicon esculentum] [Solanum lycopersicum (Lycopersicon esculentum)]	49.3	140	1.30e-41	13
gid|775|RecName: Full = Serine carboxypeptidase 2; AltName: Full = Serine carboxypeptidase II; AltName: Full = Carboxypeptidase D; AltName: Full = CPDW-II; Short = CP-WII; Contains: RecName: Full = Serine carboxypeptidase 2 chain A; AltName: Full = Serine carboxypeptidase II c [*Triticum aestivum*]	49.2	195	2.20e-55	15
gid|1542|peroxiredoxin [*Triticum aestivum*]	49.1	216	1.90e-60	4
gid|1544|troponin C [Tyrophagus putrescentiae]	49	147	8.60e-24	22
gid|650|allergen [Malassezia sympodialis]	48.9	131	4.60e-33	16
gid|1338|ragweed homologue of Art v 1 precursor [Ambrosia artemisiifolia]	48.8	84	2.10e-09	21
gid|951|Der f Mal f 6 allergen [Dermatophagoides farinae]	48.7	160	1.60e-27	20
gid|65|aldehyde dehydrogenase (NAD+) [*Alternaria alternata*]	46.3	480	1.10e-107	19
gid|64|Allergen Alt a 7 [*Alternaria alternata*]	45.7	138	1.00e-27	9
gid|2371|seed maturation-like protein precursor [Sesamum indicum]	44.5	330	3.10e-50	15
gid|2551|Par h I precursor [Parthenium hysterophorus]	44.4	81	2.00e-07	18
gid|18|Actinidain protease-like [Actinidia deliciosa]	43.8	356	9.40e-59	19
gid|775|serine carboxypeptidase II [Triticum aesivum]	43.8	153	2.70e-33	10
gid|647|allergen [Malassezia sympodialis ATCC 42132]	42.7	82	3.30e-14	3
gid|154|LMM glutenin 3 [*Triticum aestivum*]	42.5	167	6.40e-09	17
gid|1206|Sal k 3 pollen allergen [Salsola kali]	42.3	769	6.00e-94	15
gid|496|ferritin heavy chain-like protein [Dermatophagoides farinae]	42.1	183	3.90e-22	19
	42.1	164		8

(*continued on next page*)

Table 4 (*continued*)

AllergenOnline Version 18B	Highest %Seq_id	Align length	E-score	# of 23 species with matches >35% ID over 80AA
gid|496|ferritin [Dermatophagoides farinae]			4.40e-15	
gid|151|Alpha/beta gliadin-like protein product [*Triticum aestivum*]	41.8	134	1.10e-07	20
gid|150|omega-5 gliadin [*Triticum aestivum*]	41.7	396	3.00e-21	21
gid|322|beta-xylosidase [Aspergillus niger]	41.7	132	2.70e-22	11
gid|333|Taka-amylase A (Taa-G1) precursor [Aspergillus oryzae]	41.7	103	2.00e-12	1
gid|588|prohevein [Hevea brasiliensis]	41.3	121	3.00e-18	11
gid|1565|collagen alpha-2(I) chain precursor [*Bos taurus*]	41.2	131	1.70e-07	19
gid|244|Pen c 1; alkaline serine protease [Penicillium citrinum]	41.2	250	2.90e-39	1
gid|154|LMW glutenin-like protein product [*Triticum aestivum*]	40.9	235	7.70e-07	19
gid|325|PPIase [Aspergillus fumigatus]	40.8	130	5.30e-17	19
gid|588|hevein [Hevea brasiliensis]	40.8	98	3.40e-19	11
gid|63|Protein disulfide-isomerase (PDI) (Allergen Alt a 4) [*Alternaria alternata*]	40.7	81	8.70e-10	16
gid|322|xylosidase [Aspergillus niger]	40.6	256	1.20e-36	12
gid|357|trypsin [Blomia tropicalis]	40.5	237	1.00e-25	7
gid|850|catalase [Penicillium citrinum]	40.4	483	2.60e-42	21
gid|2027|allergen [Malassezia sympodialis ATCC 42132]	39.7	816	1.10e-68	21
gid|876|thioredoxin [Aspergillus fumigatus]	39.6	91	2.40e-14	16
gid|243|allergen Pen n 18 [Penicillium chrysogenum]	39.1	266	6.10e-36	2
gid|2709|lysosomal aspartic protease [*Aedes aegypti*]	38.9	522	7.30e-71	19
gid|330|manganese superoxide dismutase [Aspergillus fumigatus]	38.9	208	3.00e-29	4
gid|150|D-type LMW glutenin subunit [*Triticum aestivum*]	38.6	176	8.00e-07	21
gid|2278|thioredoxin h [*Triticum aestivum*]	38.4	86	1.20e-11	19
gid|2080|glutathione transferase [*Triticum aestivum*]	38.4	159	3.00e-15	15
gid|162|27K protein [*Triticum aestivum*]	38.2	186	4.00e-30	17
gid|785|Bromelain precursor (Allergen Ana c 2) [Ananas comosus]	38.2	152	2.70e-23	17
gid|1776|thioredoxin [Plodia interpunctella]	38.1	97	3.00e-12	19
gid|150|omega-gliadin, partial [*Triticum aestivum*]	37.7	408	9.90e-11	18
gid|833|vacuolar serine protease [Rhodotorula mucilaginosa]	37.7	297	2.80e-33	3
gid|1171|subtilisin precursor [Bacillus licheniformis]	37.6	282	2.50e-14	2
gid|18|actinidin_ [Actinidia_deliciosa]	37.5	307	2.1e-28	18
gid|160|glutenin [*Triticum aestivum*]	37.4	123	5.10e-07	10
gid|151|Gliadin-like protein product [*Triticum aestivum*]	37.1	170	8.10e-07	21

Table 4 (*continued*)

AllergenOnline Version 18B	Highest %Seq_id	Align length	E-score	# of 23 species with matches >35% ID over 80AA
gid|789|art v 2 allergen [Artemisia vulgaris]	37.1	140	5.10e-09	7
gid|1175|prepro AprM [Bacillus sp.]	37	146	7.20e-12	1
gid|875|calcium-binding protein [Ambrosia artemisiifolia]	36.7	139	4.00e-12	16
gid|987|allergen Bla g 6.0301 [*Blattella germanica*]	36.6	101	2.20e-08	12
gid|853|MPA3 allergen [*Periplaneta americana*]	36.6	243	6.60e-09	7
gid|355|cysteine protease precursor [Blomia tropicalis]	36.4	129	4.00e-12	3
gid|152|gamma-gliadin [*Triticum aestivum*]	36.3	204	1.50e-07	19
gid|793|thioredoxin [Aspergillus fumigatus]	36	86	1.70e-12	14
gid|962|putative Cup a 4 allergen [Hesperocyparis arizonica]	36	139	1.90e-09	14
gid|276|Venom allergen 5 (Antigen 5) (Ag5) (Allergen Pol f 5) (Pol f V) [Polistes fuscatus]	35.8	123	3.50e-11	1
gid|1171|RecName: Full = Subtilisin Carlsberg; Flags: Precursor [Bacillus licheniformis]	35.6	264	9.40e-09	2
gid|2576|enamine/imine deaminase [Dermatophagoides farinae]	35.5	124	4.80e-23	21
gid|151|alpha-type gliadin precursor protein [*Triticum aestivum*]	35.5	290	1.80e-07	14
gid|1174|RecName: Full = Subtilisin Savinase; AltName: Full = Alkaline protease [Bacillus lentus]	35.4	164	6.70e-20	3
gid|1959|enolase [*Salmo salar*]	35.3	428	7.60e-20	15
gid|1743|troponin C [Crangon crangon]	35.2	145	3.00e-10	16
gid|2335|chymotrypsin-like protein [*Blattella germanica*]	35.1	265	1.30e-17	7

smaller than 1e-20 to 1.1e-25. LTPs from a variety of sources have evidence of cross-reactive laboratory IgE binding, but there are fewer reports of multiple allergic reactions to diverse sources of LTPs. This search identified many proteins that are unlikely to represent major risks of cross-reactivity as the protein sequences are conserved across broad taxonomic categories with no history of cross-reactivity.

3.5. Evaluation of the limits of CODEX guidelines looking for matches of >35% identity

3.5.1. Identification of known allergens in AllergenOnline.org database using FASTA at specific E-score limits for significance

The predicted proteins from some allergenic species were compared to AllergenOnline.org database at different *E-scores*, and we focused on the best *E*-score threshold for identification of known allergens using the official WHO/IUIS Allergen Nomenclature in AOL database. Table 6 illustrates the identified allergens using FASTA in different allergenic species at representative *E*-scores of 1e-7, 1e-30, and 1e-100. All known allergens of major and minor allergenic sources in the AllergenOnline.org database were detected using *E*-scores of 10, 1, 0.001, and 1e-7. However, some potentially important matches to allergens were missed in the FASTA searches when the *E*-score was reduced less than 10e-7.

Table 5
FASTA comparison of all proteins predicted for the *Galdieria* sp. genome compared to AllergenOnline.org version 18B with >35% identity with an *E*-score of 10 or smaller. The highest percent identity matches are shown with alignment lengths and with the smallest *E* scores in columns two to four. The right-hand column shows the number of 23 common species with genome predictions that have an identity score over 35% identity to the allergens shown in the left column.

AllergenOnline Version 18B	Highest % Seq_id	Align length	E-score	# of species from the 23 genomes with >35% ID over 80AA
gid|2591|Putative heat shock-like protein [Tyrophagus putrescentiae]	72.3	653	1.00e-207	22
gid|2291|Putative Der f 33-like protein [Dermatophagoides pteronyssinus]	70.8	452	8.20e-150	23
gid|338|Putative 60S ribosomal protein L3 (Allergen Asp f 23)	69.4	385	5.80e-124	22
gid|2301|Putative glyceraldehyde-3-phosphate dehydrogenase [*Triticum aestivum*]	68.9	315	4.00e-92	21
gid|863|Putative cyclophilin [Aspergillus fumigatus]	65.5	145	5.70e-40	18
gid|1033|Allergen cytochrome *c* [Curvularia lunata]	63.1	103	1.70e-26	0
gid|2708|Putative heat shock cognate 70 [*Aedes aegypti*]	62.3	657	2.00e-172	22
gid|1959|Allergen enolase [*Salmo salar*]	62	437	2.10e-112	15
gid|1941|Putative cyclophilin [Daucus carota]	59.2	169	8.00e-41	20
gid|166|Putative triosephosphat-isomerase [*Triticum aestivum*]	59	249	1.30e65	14
gid|2236|Putative transaldolase [Cladosporium cladosporioides]	59	317	2.90e-73	6
gid|1707|Allergen aldolase A [*Thunnus albacares*]	58.9	358	3.00e-84	19
gid|651|Putative allergen [Malassezia sympodialis]	58.8	102	1.90e-24	20
gid|509|Putative 98 kDa HDM allergen [Dermatophagoides farinae]	56.3	87	6.40e-12	20
gid|1092|Putative manganese superoxide dismutase-like protein [Pistacia vera]	55.1	207	8.30e-52	17
gid|62|Putative RecName: Full = 60S acidic ribosomal protein P2; AltName: Full = Minor allergen Alt a 5; AltName: Full = Allergen Alt a 6; AltName: Full = Allergen Alt a VI; AltName: Allergen = Alt a 5	54.8	115	1.30e-19	10
gid|1983|Putative 60S acidic ribosomal phosphoprotein P1 [Penicillium crustosum]	52.7	112	2.20e-22	16
gid|64|Putative Minor allergen Alt a 7 (Alt a VII)	52	202	9.40e-39	15
	50	106	2.00e-18	0

Table 5 (*continued*)

AllergenOnline Version 18B	Highest % Seq_id	Align length	E-score	# of species from the 23 genomes with >35% ID over 80AA
gid|1026|Allergen allergen [Malassezia sympodialis ATCC 42132]				
gid|325|Allergen PPIase [Aspergillus fumigatus]	48.9	135	2.80e-19	19
gid|1926|Allergen cyclophilin [Catharanthus roseus]	47.6	170	5.80e-32	18
gid|1544|Putative troponin C [Tyrophagus putrescentiae]	45.9	146	3.80e-27	22
gid|1248|Putative eukaryotic translation initiation factor [Forcipomyia taiwana]	45.2	325	4.10e-64	21
gid|1206|Allergen Sal k 3 pollen allergen [Salsola kali]	45.1	765	8.60e-77	15
gid|2849|Allergen Chain A, Beta-amylase	44	470	1.80e-59	0
gid|2582|Putative alcohol dehydrogenase [Curvularia lunata]	43.4	339	1.00e-61	8
gid|951|Allergen Der f Mal f 6 allergen [Dermatophagoides farinae]	43.4	143	2.30e-19	20
gid|496|Allergen ferritin heavy chain-like protein [Dermatophagoides pteronyssinus]	42.5	179	2.60e-25	19
gid|63|Putative Protein disulfide-isomerase (PDI) (Allergen Alt a 4)	42.4	92	7.50e-10	16
gid|65|Putative aldehyde dehydrogenase (NAD+) [*Alternaria alternata*]	42.3	506	5.60e-75	19
gid|246|Putative elongation factor 1 beta-like [Penicillium citrinum]	42.1	235	3.30e-36	20
gid|2076|Putative heat shock protein 70 [Dermatophagoides farinae]	40.4	560	2.60e-61	10
gid|850|Putative catalase [Penicillium citrinum]	39.9	489	4.90e-71	21
gid|2293|Allergen Der f 31 allergen [Dermatophagoides farinae]	39.6	144	1.90e-12	11
gid|1617|Putative alpha/beta gliadin precursor [*Triticum aestivum*]	39.1	161	1.10e-12	13
gid|2592|Putative aldehyde dehydrogenase-like protein [Tyrophagus putrescentiae]	38.5	405	3.90e-59	20
gid|251|Putative peroxisomal membrane protein [Penicillium citrinum]	37.8	172	2.50e-16	2
gid|650|Putative allergen [Malassezia sympodialis]	37.5	144	7.50e-22	16
gid|160|Allergen high molecular weight glutenin subunit 1Ax1 [*Triticum aestivum*]	36.4	110	3.30e-07	4
gid|2215|Allergen RecName: Full = Glutathione S-transferase 1; AltName: Full = GST class-sigma	36.3	204	1.60e-19	4

(*continued on next page*)

Table 5 (*continued*)

AllergenOnline Version 18B	Highest % Seq_id	Align length	E-score	# of species from the 23 genomes with >35% ID over 80AA
gid|799|Allergen NADP-dependent mannitol dehydrogenase [Davidiella tassiana]	36.2	246	4.20e-24	5
gid|1577|Allergen Sal k 4.03 allergen [Salsola kali]	35.8	148	6.10e-12	0
gid|2576|Putative enamine/imine deaminase [Dermatophagoides farinae]	35.7	126	2.40e-12	21
gid|1171|Allergen subtilisin precursor [Bacillus licheniformis]	35.4	178	4.20e-14	2
gid|2551|Putative Par h I precursor [Parthenium hysterophorus]	35.2	145	9.60e-12	18

3.5.2. Major allergens with a high risk of cross-reactivity

Proteins predicted from the 23 genomes that included humans were searched to allergens having a relatively high risk of clinical cross-reactivity to major allergens. The distribution of taxa having matches to clinically important major allergens included lipid transfer proteins, vicilins, glycinins, 2S albumins, tropomyosin, and arginine kinase are shown in Table 7. The matches were related to taxonomic relationships of the species as well as the protein families, yet the identity matches are broadly diverse. Lipid transfer proteins, vicillins and glycinins are highly conserved in beans, soybeans, apple, peach, and papaya. Yet publications of cross-reactivity for these proteins among the protein families is limited and it appears true clinical cross-reactivity is very limited. Major allergens in crustacean shellfish e.g. tropomyosin and arginine kinase are generally cross-reactive between crustaceans and occasionally to insect proteins, yet identities of >35% identity were commonly found for those two proteins between human, drosophila, bovine, salmon, and cod and there is no evidence of clinical cross-reactivity for those taxa compared to crustaceans. Importantly, human proteins are not considered to be allergenic for humans.

3.5.3. Minor allergens and noise of CODEX limits

To consider protein identity matches to minor allergens, predicted proteins of the 23 species were compared to AOL version 18B by FASTA using the HCC supercomputer. Those proteins that had a match of >35% identity to proteins from at least 10 of these species were considered evolutionarily conserved minor allergens. Most of the minor allergens represented had sequence identities less than 50% when compared within the protein type. Matches of >35% identity were found to 170 allergens listed in AOL, and those are considered minor also because they do not have published evidence of causing clinical reactions, only IgE binding. They all matched at least 10 different species out of the 23

Table 6
Identification of known allergens listed in the AllergenOnline.org database using full-length FASTA. Predicted proteins of the allergenic species listed in this table were compared to the AOL database. Matches above CODEX limits to known allergens were found using the official WHO/IUIS Allergen Nomenclature in the AOL database identified *E*-scores from 1e-7, 1e-30 and 1e-100. A few allergens were missed between *E*-scores of 1e-7 and 1e-30.

Species	E-score (1e-7)	E-score (1e-30)	E-score (1e-100)
Peanut (*Arachis hypogaea*)	Ara h 1, Ara h 2, Ara h 3, Ara h 4, Ara h 6, Ara h 7, Ara h 8, profilin, lipid transfer proteins, oleosin, conarachin, glycinin	Ara h 1, Ara h 2, Ara h 3, Ara h 4, Ara h 6, Ara h 7, Ara h 8, profilin, lipid transfer proteins, oleosin, conarachin, glycinin	Ara h 1, Ara h 3, Ara h 4, conarachin, glycinin
Apple (*Malus domestica*)	Mal d 1, Mal d 2, Mal d 3, Mal d 4	Mal d 1, Mal d 2, Mal d 3, Mal d 4	Mal d 2
Chicken (*Gallus gallus*)	Gal d 1, Gal d 2, Gal d 3, Gal d 4, Gal d 5, Gal d 7, Gal d 8	Gal d 1, Gal d 2, Gal d 3, Gal d 4, Gal d 5, Gal d 7, Gal d 8,	Gal d 1, Gal d 2, Gal d 3
Soybeans (*Glycine max*)	Gly m 1, Gly m 3, Gly m 5, Gly m 6, Gly m 8, Gly m Bd 28K, Gly m Bd 30K, glycine trypsin inhibitor	Gly m 3, Gly m 5, Gly m 6, Gly m 8, Gly m Bd 28K, Gly m Bd 30K, glycine trypsin inhibitor	Gly m 5, Gly m 6
Bovine (*Bos taurus*)	Bos d 2, Bos d 3, Bos d 4, Bos d 5, Bos d 6, Bos d 9 or Bos d 10, Bos d 11, Bos d 12	Bos d 2, Bos d 3, Bos d 4, Bos d 5, Bos d 6, Bos d 11, Bos d 12	Bos d 6
Candida (*Candida albicans*)	Cand a 1, Cand a 3, Enolase 1	Cand a 1, Cand a 3, Enolase 1	Cand a 1, Cand a 3, Enolase 1
Cod (*Gadus morhua*)	Gad m 1	Gad m 1	
Papaya (*Carica papaya*)	Cari p 1.0101	Cari p 1.0101	Cari p 1.0101
Almond (*Prunus dulcis*)	Pru du 1.01, Pru du 2, Pru 4, Pru du 6	Pru du 1.01, Pru du 2, Pru 4 Pru du 6	Pru du 2, Pru du 6
Rice (*Oryza sativa*)	Glyoxalase I, Ory s 1, Polcalcin (Ph1p7)	Glyoxalase I	Glyoxalase I
Pecan (*Carya illinoinensis*)	Car i 1, Car i 4	Car i 4	Car i 4
Bean (*Phaseolus vulgaris*)	Pha v 3	Pha v 3	
Pistacio (*Pistacia vera*)	Pis v 1, Pis v 2 (2.0201), Pis v 3, Pis v 4, Pis v 5	Pis v 2 (2.0201), Pis v 3, Pis v 4, Pis v 5	Pis v 2 (2.0201), Pis v 3, Pis v 4, Pis v 5
Peach (*Prunus persica*)	Pru p 1 and 1.0201, Pru p 2, 2.01A, 2.01B, 2.02, Pru p 3, Pru du 4.02	Pru p 1 and 1.0201, Pru p 2 (2.01A, 2.01B, 2.02), Pru p 3, Pru du 4.02	
Salmon (*Salmo salar*)	Sal s 1, Sal s 2, Sal s 3	Sal s 1, Sal s 2, Sal s 3	Sal s 1, Sal s 2, Sal s 3
Potato (*Solanum tuberosum*)	Sola t 1, Sola t 2, Sola t 3, Sola t 4, profilin	Sola t 1, Sola t 2, Sola t 3, Sola t 4 profilin,	Sola t 1
Walnut (*Juglan regia*)	Jug r 1, Jug r 2, Jug r 3	Jug r 2, Jug r 3	
Wheat (*Triticum aestivum*)	Tri a 12, Tri a 14, Tri a 21, Tri a 25, Tri a 26, Tri a 28, Tri a 29, Tri a 31, Tri a 33, Tri a 34, Tri a 37, Tri a 39, Tri a 42, Tri a 44, Tri a 45, thaumatin like protein, serine carboxypeptidase, serine carboxypeptidase, putative 27K protein, chymotrypsin inhibitor WSCI	Tri a 12, Tri a 14, Tri a 21, Tri a 25, Tri a 26, Tri a 28, Tri a 29, Tri a 31, Tri a 33, Tri a 34, Tri a 37, Tri a 39, Tri a 42, Tri a 44, Tri a 45, thaumatin like protein, serine carboxypeptidase, putative 27K protein, chymotrypsin inhibitor WSCI	Tri a 31, Tri a 33, Tri a 34, serine carboxypeptidase

Table 7
Distribution of matches of proteins predicted from the 23 species genomes to clinically important allergens in AllergenOnline.org. The matches were identified based on CODEX guidelines of >35% identity over 80 AA and would be considered as possibly cross-reactive yet, human proteins matched in this search are clearly not allergens, demonstrating over-prediction.

LTPs	Vicilins	Glycinins	Tropomyosins	Arginine kinase	2S albumins
Peanut	Papaya	Soybeans	Drosophila	Human	Pistachio
Kidney beans	Corn	Kidney beans	Salmon	Chicken	Potato
Walnut	Drosophila	Peanut	Atlantic cod	Bovine	Soybeans
Soybeans	Pistachio	Salmon	Chicken	Salmon	Walnut
Apple	Soybeans	Walnut	Human	Atlantic cod	Peanut
Papaya	Peanut	Chicken	Bovine	Drosophila	
Rice	Almond	Human			
Wheat	Pecan	Potato			
Peach	Walnut				
Potato	Potato				
Bovine	Apple				
Human	Peach				
Corn	Human				
Arabidopsis	Salmon				
Almond					

species. Supplementary Table 2 lists these minor allergens and shows the number of species matched out of 23. Searches that identify protein identity matches to proteins in 10 or more diverse species must be evolutionarily conserved and are unlikely to represent real risks of cross-reactivity.

4. Conclusion

It is becoming more common to use a whole genome or a proteome bioinformatics approach to identify potential proteins in a wide variety of species. Some regulatory agencies or risk assessment scientists have suggested using these predicted proteins against allergen databases to identify possible risks of allergenicity for food safety. The CODEX guideline (>35% identity over 80 amino acids to any known allergen) has become a standard for possible risks of cross-reactivity since 2003. The comparison to www.AllergenOnline.org was made available to the public in 2005 to assess individual proteins. The database is updated annually. The interpretation of identity matches over 35% over 80 amino acids or the equivalent is assumed to be a positive identity match that would require serum IgE binding tests sera from subjects allergic to the matched allergen. Since we know that matches at that identity level can occur by random chance, we tested the use of protein sequences predicted from genomes, transcriptomes, or proteomes against AOL to estimate the commonality of false positive matches.

We compared the predicted proteins from the genomes of 23 diverse allergenic and low- or non-allergenic species including plant sources, fungi, fish, insect and other animal sources as well as human sequences against the AOL database using standard CODEX criteria as well as full-FASTA alignments to provide identity matches. We used a wide variety of *E* score criteria to consider that as a variable as well. A number of housekeeping proteins across many species had moderate to high identities to minor putative allergens in AOL. However, many of these proteins are highly conserved in most eukaryotes and as a consequence would be expected to be found in any search using the standard CODEX criteria. In contrast, major allergens are not highly conserved in sequence and structure and were not identified using the search parameters except in closely related species.

For those highly conserved proteins identified across many species, there are nonetheless differences in the levels of AA sequence identity conservation that impact their potential for shared clinical cross-reactivity. Moreover, differences in protein abundance and potency are significantly different between species, affecting the allergenic potential of the species.

We have used a wide range of *E*-score thresholds to test search methods. We propose that an *E*-score threshold of 1e-7 may be needed for identification of a few important allergens in this type of study, yet identity matches of >35% are still common for highly conserved proteins at 1e-7.

Examples using three predicted proteomes from three novel foods were assessed against the AOL database and many identity matches were seen. The comparison of predicted proteins from 23 test species demonstrated conclusively that the low-level match of >35% identity over 80 amino acids over-predicts potential risks of allergy. We have concluded that *Chlorella variabilis*, *Galdieria sulphuraria* and *Fusarium strain flavolapis* do not represent a significant risk of food allergy to the general population as matches to similar proteins from many diverse species are very common.

Alternative strategies of increasing the match criteria above 35% identity, possibly to 45% identity; decreasing the *E*-score below 1e-7 or smaller; may be needed although matches to a few allergens may be missed at 1e-20 and ranking of allergens in AOL regarding risks of disease could markedly improve this assessment strategy. Other investigators should use similar strategies and risk assessors should consider the broad questions of whole food safety for novel or new foods to establish more predictive assessment limits.

Author contributions

MA performed the initial literature reviews, performed bioinformatics comparisons to AllergenOnline.org, and drafted the manuscript. CZ oversaw the design of the bioinformatics pipeline and edited the manuscript. MK oversaw generation of the genomic sequences of *Fusarium* strain flavolapis and BF contributed to allergenicity discussion of and edits *Fusarium* strain flavolapis. MC oversaw and provided genomic data for *Galdieria sulphuraria*. HG provided overall suggestions and edited the manuscript. REG developed the original concept for the study and oversaw the completion and provided allergenicity risk review.

Funding

Some funds were provided by Fermentalg and by Nature's Fynd Some funding was provided by the AllergenOnline.org sponsors (Unilever and NuSeed). The majority was from revolving research accounts of Professor Goodman.

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.

Acknowledgements

Genomic sequences of *Galdieria* sp. were provided by Fermentalg and for *Fusarium* sp. by Nature's Fynd. The AllergenOnline.org database search routines were developed by John Wise as an employee of the University of Nebraska. Jean-Jack Riethoven of the Holland Computing Center at the University of Nebraska provided BLASTP supercomputer access for much of this work and loaded the NCBI non-redundant Protein dataset on HCC.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.fct.2020.111888.

References

Aalberse, R.C., Kleine Budde, I., Stapel, S.O., van Ree, R., 2001. Structural aspects of cross-reactivity and its relation to antibody affinity. Allergy 56 (Suppl. 67), 27–29.
Aalberse, R.C., 2000. Structural biology of allergens. J. Allergy Clin. Immunol. 106, 228–238.
Andrews, S., 2010. FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc.
Bankevich, A., Nurk, S., Antipov, D., et al., 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. 19, 455–477.
Blanc, G., Duncan, G., Agarkova, I., Borodovsky, M., Gurnon, J., Kuo, A., Lindquist, E., Lucas, S., Pangilinan, J., Polle, J., et al., 2010. The Chlorella variabilis NC64A genome reveals adaptation to photosymbiosis, coevolution with viruses, and cryptic sex. Plant Cell 22, 2943–2955.
Bleakley, S., Hayes, M., 2017. Algal proteins: extraction, application, and challenges concerning production. Foods 6, 33.
Cabanillas, B., Jappe, U., Novak, N., 2018. Allergy to peanut, soybean and other legumes: recent advances in allergen characterization, stability to processing and IgE cross-reactivity. Mol. Nutr. Food Res. 62, 1.
CODEX Alimentarius Commission, 2009. Foods Derived from Modern Biotechnology, second ed. World Health Organization, Food and Agricultural Organization of the United Nations, Rome, Italy.
Cressman, R.F., Ladics, G., 2009. Further evaluation of the utility of "sliding window" FASTA in predicting cross-reactivity with allergenic proteins. Regul. Toxicol. Pharmacol. 54 (3 Suppl. l), S20–S25.
Dreskin, S.C., Germinaro, M., Reinhold, D., et al., 2019. IgE binding to linear epitopes of Ara h 2 in peanut allergic preschool children undergoing oral Immunotherapy. Pediatr. Allergy Immunol. 30, 817–823.
FAO/WHO, 2001. Evaluation of allergenicity of genetically modified foods. In: Report of a Joint FAO/WHO Expert Consultation on Allergenicity of Foods Derived from Biotechnology. Food and Agriculture Organization of the United Nations (FAO), Rome.
Faber, M.A., Pascal, M., El Karbouchi, O., Sabato, V., Hagendorens, M.M., Decuyper, I.I., Bridts, C.H., Ebo, D.G., 2017. Shellfish allergens: tropomyosin and beyond. Allergy 72, 842–848.
Finnigan, T.J.A., Wall, B.T., Wilde, P.J., Stephens, F.B., Taylor, S.L., Freedman, M.R., 2019. Mycoprotein: the future of nutritious nonmeat protein, a symposium review. Curr. Dev. Nutr. 3, nzz021.
Frigerio, J., Agostinetto, G., Sandionigi, A., Mezzasalma, V., Maria Berterame, N., Casiraghi, M., Labra, M., Galimberti, A., 2020. The hidden 'plant side' of insect novel foods: a DNA-based assessment. Food Res. Int. 128, 108751.
Goodman, R.E., Vieths, S., Sampson, H.A., Hill, D., Ebisawa, M., Taylor, S.L., van Ree, R., 2008. Allergenicity assessment of genetically modified crops—what makes sense? Nat. Biotechnol. 26 (1), 73–81. https://doi.org/10.1038/nbt1343.
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.
Goodman, R.E., Hefle, S.L., Taylor, S.L., van Ree, R., 2005. Assessing genetically modified crops to minimize the risk of increased food allergy: a review. Int. Arch. Allergy Immunol. 137, 153–166.
Gowland, M.H., Walker, M.J., 2015. Food allergy, a summary of eight cases in the UK criminal and civil courts: effective last resort for vulnerable consumers? J. Sci. Food Agric. 95, 1979–1990.
Gurevich, A., Saveliev, V., Vyahhi, N., Tesler, G., 2013. QUAST: quality assessment tool for genome assemblies. Bioinformatics 29, 1072–1075.
Henikoff, J.G., Henikoff, S., 1996. Blocks database and its applications. Methods Enzymol. 266, 88–105.
Henikoff, S., Henikoff, J.G., 1992. Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. U.S.A. 89, 10915–10919.
Hileman, R.E., Silvanovich, A., Goodman, R.E., Rice, E.A., Holleschak, G., Astwood, J.D., Hefle, S.L., 2002. Bioinformatic methods for allergenicity assessment using a comprehensive allergen database. Int. Arch. Allergy Immunol. 128, 280–291.
Hoff, M., Ballmer-Weber, B.K., Niggemann, B., Cistero-Bahima, A., San Miguel-Moncin, M., Conti, A., Haustein, D., Vieths, S., 2003b. Molecular cloning and immunological characterization of potential allergens from the mould Fusarium culmorum. Mol. Immunol. 39, 965–975.
Hoff, M., Trueb, R.M., Ballmer-Weber, B.K., Vieths, S., Wuethrich, B., 2003a. Immediatetype hypersensitivity reaction to ingestion of mycoprotein (Quorn) in a patient allergic to molds caused by acidic ribosomal protein P2. J. Allergy Clin. Immunol. 111, 1106–1110.
Hyatt, D., Chen, G.L., Locascio, P.F., Land, M.L., Larimer, F.W., Hauser, L.J., 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinf. 11, 119.
Jacobson, M.F., DePorter, J., 2018. Self-reported adverse reactions associated with mycoprotein (Quorn-brand) containing foods. Ann. Allergy Asthma Immunol. 120, 626–630.
Katona, S.J., Kaminski, E.R., 2002. Sensitivity to Quorn mycoprotein (Fusarium venenatum) in a mould allergic patient. J. Clin. Pathol. 55, 87–88.
King, R., Urban, M., Hammond-Kosack, M.C.U., Hassani-Pak, K., Hammond-Kosak, K.E. H., 2015. The completed genome sequence of the pathogenic ascomycete fungus Fusarium graminearum. BMC Genom. 16, 544.
Klamczynska, B., Mooney, W.D., 2017. Heterotrophic Microalgae: a Scalable and Sustainable Protein Source. Sustainable Protein Sources.
Ladics, G.S., Bannon, G.A., Silvanovich, A., Cressman, R.F., 2007. Comparison of conventional FASTA identity searches with the 80 amino acid sliding window FASTA search for the elucidation of potential identities to known allergens. Mol. Nutr. Food Res. 51, 985–998.
Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359.
Li, H., 2016. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. Bioinformatics 32, 2103–2110.
Li, H., Durbin, R., 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25, 1754–1760.
Lowe, T.M., Chan, P.P., 2016. tRNAscan-SE On-Line: search and contextural analysis of Transfer RNA genes. Nucleic Acids Res. 44, W54–W57.
Lowe, T.M., Eddy, S.R., 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 25 (5), 955–964.
Metcalfe, D.D., Astwood, J.D., Townsend, R., Sampson, H.A., Taylor, S.L., Fuchs, R.L., 1996. Assessment of the allergenic potential of foods derived from genetically engineered crop plants. Crit. Rev. Food Sci. Nutr. 36 (Suppl. ment), S165–S186.
Muraro, A., Hoffmann-Sommergruber, K., Holzhauser, T., Poulsen, L.K., Gowland, M.H., Akdis, C.A., Mills, E.N.C., et al., 2014. EAACI food allergy and anaphylaxis guidelines. Protecting consumers with food allergies: Allergy 69, 1464–1472.
Niehaus, E.-M., Munsterkotter, M., Proctor, R.H., Brown, D.W., Sharon, A., Idan, Y., Oren-Young, L., et al., 2016. Comparative "Omics" of the Fusarium fujikuroi species complex highlights differences in genetic potential and metabolite synthesis. Genome Biol. Evol. 8 (11), 3574–3599.
Palladino, C., Breiteneder, H., 2018. Peanut allergens. Mol. Immunol. 100, 58–70.
Pearson, W.R., 2000. Flexible sequence similarity searching with the FASTA3 program package. Methods Mol. Biol. 132, 185–219.
Pearson, W.R., 2014. BLAST and FASTA similarity searching for multiple sequence alignment. Methods Mol. Biol. 1079, 75–101.
Pearson, W.R., 2016. Finding protein and nucleotide similarities with FASTA. Curr. Protoc. Bioinformatics 53, 3.9.1-25.
Porterfield, H.S., Murray, K.S., Schlichting, D.G., Chen, X., Hansen, K.C., Duncan, M.W., Dreskin, S.C., 2009. Effector activity of peanut allergens: a critical role for Ara h 2, Ara h 6, and their variants. Clin. Exp. Allergy 39, 1099–1108.
Ramsey, N.B., Duffey, D., Anagnostou, K., Coleman, N.E., Davis, C.M., 2019. Epidemiology of anaphylaxis in critically ill children in the United States and Canada. J. Allergy Clin. Immunol. Pract. 7, 2241–2249.
Ruethers, T., Taki, A.C., Johnston, E.B., Nugraha, R., Le, T.T.K., Kalic, T., McLen, T.R., Kamath, S.D., Lopata, A.L., 2018. Seafood allergy: a comprehensive review of fish and shellfish allergens. Mol. Immunol. 100, 28–57.
Schonknecht, G., Chen, W.-H., Ternes, C.M., Barbier, G.G., Shrestha, R.P., Stanke, M., Brautigam, A., Baker, B.J., Banfield, J.F., Garavito, R.M., et al., 2013. Gene transfer from Bacteria and Archaea facilitated evolution of an extremophilic eukaryote. Science 339, 1207–1210.
Schwager, C., Kull, S., Behrends, J., Rockendorf, N., Schocker, F., Frey, A., Homann, A., Becker, W.-.M., Jappe, U., 2017. Peanut oleosins associated with severe peanut allergy—importance of lipophilic allergens for comprehensive allergy diagnostics. J. Allergy Clin. Immunol. 140, 1331–1338.
Silvanovich, A., Bannon, G., McClain, S., 2009. The use of E-scores to determine the quality of protein alignments. Regul. Toxicol. Pharmacol. 54 (3 Suppl. l), S26–S31.
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.
Stanke, M., Morgenstern, B., 2005. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. Nucleic Acids Res. 33 (Web Server issue), W465–W467.
Taylor, S.L., Hefle, S.L., 2006. Food allergen labeling in the USA and Europe. Curr. Opin. Allergy Clin. Immunol. 6 (3), 186–190.
Tee, R.D., Gordon, D.J., Welch, J.A., Newman Taylor, A.J., 1993. Investigation of possible adverse allergic reactions to mycoprotein ('Quorn'). Clin. Exp. Allergy 23, 257–260.
Thomas, K., Bannon, G., Hefle, S., Herouet, C., Holsapple, M., Ladics, G., Macintosh, S., Privalle, L., 2005. In silico methods for evaluating human allergenicity to novel proteins: international bioinformatics workshop meeting report, 23-24 February. Toxicol. Sci. 88, 307–310.

Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., et al., 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PloS One 9, e112963.
Weber, R.W., Levetin, E., 2014. Allergen of the month-Fusarium. Ann. Allergy Asthma Immunol. 112, A11.
Wells, M.L., Potin, P., Craigie, J.S., Raven, J.A., et al., 2017. Algae as nutritional and functional food sources: revisiting our understanding. J. Appl. Phycol. 29, 949–982.
Westerhout, J., Baumert, J.L., Blom, W.M., Allen, K.J., et al., 2019. Deriving individual threshold doses from clinical challenge data for population risk assessment of food allergens. J. Allergy Clin. Immunol. 144, 1290–1309.
Wilder, H.K., Raffel, S.J., Barbour, A.G., Porcella, S.F., Sturdevant, D.E., Vaisvil, B., Kapatral, V., Schmitt, D.P., Schwan, T.G., Lopez, J.E., 2016. Transcriptional profiling the 150 kb linear megaplasmid of *Borrelia turicatae* suggests , a role in vector colonization and initiating mammalian infection. PloS One 11 (2), 1–17.
Yeh, C.C., Tai, H.Y., Chou, H., Wu, K.G., Shen, H.D., 2016. Acuolar serine protease is a major allergen of Fusarium proliferatum and an IgE-cross reactive pan-fungal allergen. Allergy Asthma Immunol. Res. 8, 438–444.
Zimin, A.V., Marcais, G., Puiu, D., Roberts, M., Salzberg, S.L., Yorke, J.A., 2013. The MaSuRCA genome assembler. Bioinformatics 29, 2669–2677.