



Q-X1-P-X2 motif search for potential celiac disease risk has poor selectivity

Ping Song^{a,*}, Nancy Podevin^b, Henry Mirsky^b, Jennifer Anderson^b, Bryan Delaney^b,
Carey Mathesius^b, Laura Rowe^b, Rod A. Herman^a

^a Corteva Agriscience™, Agriculture Division of DowDuPont™, Indianapolis, IN, USA

^b Corteva Agriscience™, Agriculture Division of DowDuPont™, Johnston, IA, USA

ARTICLE INFO

Keywords:

Celiac disease

Bioinformatics

Motif

Selectivity

GM crops

ABSTRACT

The European Food Safety Authority (EFSA) recently published guidelines for assessment of potential celiac disease risk for newly expressed proteins in genetically modified (GM) crops. This novel step-wise approach prescribes, in part, how to conduct sequence identity searches between a newly expressed protein and known celiac disease peptides including a Q/E-X1-P-X2 amino acid motif. To evaluate the specificity of the recommended sequence identity searches in the context of risk assessment, protein sequences from celiac disease causing crops, as well as from crops not associated with celiac disease, were compared with known HLA-DQ restricted epitopes and searched for the presence of motifs followed by peptide analysis. Searches for the presence of the Q/E-X1-P-X2-motif were found to generate a high proportion of false-positive hits irrelevant to celiac disease risk. Identification of a 9mer exact match between a newly expressed protein and the known celiac disease peptides (recommended by the guideline) along with a supplementary sequence comparisons (suggested by FARRP/AllergenOnline) is considered better suited to more specifically capture the potential risk of a newly expressed protein for celiac disease.

1. Introduction

Genetically modified (GM) crops have been adopted by farmers in many countries at an unprecedented speed. In 2017, worldwide cultivation acreage of GM crops reached a record of 189.8 million hectares, a ~111-fold increase over a period of 22 years since their initial commercial launch in 1996 (ISAAA, 2017). The wide adoption of GM crop cultivation has brought economic and social benefits, and farm sustainability (National Research Council, 2010; Kathage and Qaim, 2012; Klümper and Qaim, 2014). Notably, there have been no safety incidents involving food/feed derived from GM crops, although the debate on the safety of GM crops has continued for many years. One of the most commonly voiced food safety concern is potential allergenicity of introduced proteins. To protect human food safety and public health, international organizations, including the United Nations Food and Agriculture Organization (FAO) and World Health Organization (WHO) and their joint Codex Alimentarius Commission, have issued guidelines on evaluation of allergenicity of foods derived from biotechnology (FAO/WHO, 2001; Codex Alimentarius Commission, 2009).

Despite the existence of rigorous regulations on food/feed derived through modern biotechnology and the absence of safety incidences involving food/feed derived from GM crops, certain regulatory

authorities consider that adding requirements will increase the public trust in the risk assessment process. However, such regulation may inadvertently send the message that risks are now believed to be greater than previously thought, when in fact, experience and acquired knowledge has shown negligible risk exists. EFSA (European Food Safety Authority) recently published an additional guideline on the allergenicity assessment of GM crops (Naegeli et al., 2017). According to this guideline, newly expressed proteins in GM crops must be subjected to a step-wise risk assessment for potential risk of non-IgE-mediated allergic responses with particular emphasis on differentiating them from proteins known to provoke adverse effects in persons with celiac disease (CD). Fortunately, the proteins that cause celiac disease are not widely distributed in nature being largely restricted to gluten proteins from wheat (*Triticum aestivum*), gluten-like hordeins from barley (*Hordeum vulgare*), and secalins from rye (*Secale cereal*) and derived hybrids. Oats (*Avena sativa*) are generally considered relatively safe for celiac disease patients although this crop contains gluten-like avenins (Naegeli et al., 2017). Gluten is a composite of storage proteins localized together with starch in the endosperm of various cereal grains. “Gluten” proteins causing celiac disease are limited to the grains listed above. Gluten proteins are characterized by the abundant presence of glutamine and proline which together comprise over 50% of the amino

* Corresponding author.

E-mail address: psong@dow.com (P. Song).

<https://doi.org/10.1016/j.yrtph.2018.09.021>

Received 30 July 2018; Received in revised form 19 September 2018; Accepted 21 September 2018

Available online 26 September 2018

0273-2300/ © 2018 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/>).

acids (Wieser, 2007).

In one component of the step-wise approach outlined by EFSA, the sequence identity between a newly expressed protein and celiac disease peptides must be evaluated if knowledge of the protein including source organism, exposure, etc., is insufficient to support the history of safe use of that protein. Notably, this must be done without taking into account the source organism for the gene encoding the protein and whether or not it is a known gluten-containing plant (wheat, barley, rye, and oats) (Figure 1 in Naegeli et al., 2017). Such sequence identity evaluation includes a search for a Q/E-X1-P-X2 motif (Q = glutamine; E = glutamic acid; X1 = L [leucine], Q, F [phenylalanine], S [serine], or E; P = proline; X2 = Y [tyrosine], F, A [alanine], V [valine] or Q) present in the HLA-DQ2 restricted epitopes. The restricted epitopes associated with celiac disease, HLA-DQ2 and HLA-DQ8, are nine amino acids long, and the Q/E-X1-P-X2 is present in most, but not all of the HLA-DQ2 restricted epitopes (Sollid et al., 2012). One would expect a motif consisting of only four amino acids to be found in many protein sequences that are not related to celiac disease, thus creating a large number of 9mers requiring further evaluation by HLA-DQ peptide modeling according to the stepwise approach of the risk assessment for celiac disease as outlined in the EFSA guideline.

To fully understand the performance of the sequence identity searches for identifying potential celiac disease risk, especially the Q/E-X1-P-X2 motif search, whole protein amino acid sequences (*in silico* translated from genomic DNA from various food plants and animals) available in NCBI (<ftp://ftp.ncbi.nih.gov/genomes/>) and all the protein sequences of a given food plant or animal available in the UniProtKB (<http://www.uniprot.org/>) database were downloaded and used to investigate their sequence identity to HLA-DQ2/DQ8 restricted epitopes. The outcomes were evaluated to determine the specificity of the proposed approach.

2. Material and method

2.1. Protein sequences

Protein sequences from food plants and animals including apple (*Malus domestica*), sunflower (*Helianthus annuus*), tomato (*Solanum lycopersicum*), rice (*Oryza sativa*), potato (*Solanum tuberosum*), soybean (*Glycine max*), banana (*Musa acuminata*), maize (*Zea mays*), cattle (*Bos taurus*), and swine (*Sus scrofa*) were downloaded from the NCBI genome FTP site (<ftp://ftp.ncbi.nih.gov/genomes/>; as of December 2017). These proteins were *in silico* translated from the coding sequences in the whole genome, which reflects a variety of proteins to which a consumer might be exposed. Similar protein sequences are currently unavailable for wheat, oats, barley, and rye. However, the protein sequences from the genome of red wild einkorn wheat (*Triticum Urartu*; A genome), the contributor of one of the sub-genomes of bread wheat (which is hexaploid; AABBDD) was available from a recent sequencing effort, and was used in this investigation. In addition, all the protein sequences available in the UniProtKB database for wheat, barley, oats, rye, sunflower, apple, banana, peanut (*Arachis hypogaea*), potato, swine, chicken (*Gallus gallus*), and cattle were downloaded (as of January 2018). It is noteworthy that the protein sequences downloaded from the UniProtKB database might contain redundant sequences or bias toward certain types of proteins that do not reflect the whole protein set in a given species due to the nature of this database.

2.2. Search for match of HLA-DQ2 and HLA-DQ8 restricted epitopes

The 22 HLA-DQ2 and Four HLA-DQ8 restricted epitopes (9 amino acids long) listed in the EFSA guideline (Naegeli et al., 2017) along with four HLA-DQ2 restricted epitopes downloaded from the ProPepper website (<https://propepper.net/epitope-list>) were used for sequence identity searches (Table 1). Since the Q/E-X1-P-X2 motif is not present in one of the 22 HLA-DQ2 epitopes in the list provided in the EFSA

Table 1

List of HLA-DQ restricted epitopes for sequence identity search.

(Source: Naegeli et al., 2017; ProPepper: <https://propepper.net/epitope-list>)

Epitope Name	Sequence	Search Pattern
DQ2.5-glia-α1a	PFPQPQLPY	9mer exact match
DQ2.5-glia-α1b	PYPQPQLPY	
DQ2.5-glia-α2	PQPQLPYQP	9mer match allowing three mismatches
DQ2.5-glia-α3	FRPQQPYQP	
DQ2.5-glia-γ1	PQQSFPQQQ	9mer exact match
DQ2.5-glia-γ2	IQPQQPAQL	
DQ2.5-glia-γ3	QQPQQPYQP	9mer match allowing three mismatches
DQ2.5-glia-γ4a	SQPQQQFPQ	
DQ2.5-glia-γ4b	PQPQQQFPQ	9mer match allowing three mismatches
DQ2.5-glia-γ4c	QQPQQPFPQ	
DQ2.5-glia-γ4d	PQPQQPFCQ	9mer match allowing three mismatches
DQ2.5-glia-γ5	QQPFPQQPQ	
DQ2.5-glia-ω1	PFPQPQQPF	9mer match allowing three mismatches
DQ2.5-glia-ω2	PQPQQPFPW	
DQ2.2-glut-L1	PFSQQQQPV	9mer match allowing three mismatches
DQ2.5-glut-L2	FSQQQQSPF	
DQ2.5-hor-1	PFPQPQQPF	9mer match allowing three mismatches
DQ2.5-hor-2	PQPQQPFPQ	
DQ2.5-sec-1	PFPQPQQPF	9mer match allowing three mismatches
DQ2.5-sec-2	PQPQQPFPQ	
DQ2.5-ave-1	PYPEQQEPF	9mer match allowing three mismatches
DQ2.5-ave-1b	PYPEQQQPF	
DQ2-LMW-glutenin	FSQQQQQPL	9mer match allowing three mismatches
DQ2.5-hor-2, DQ2.5-sec-2	PQPQQPFPQ	
DQ2.2-glut-L1, DQ2.5-glut-L1	PFSQQQQPV	9mer match allowing three mismatches
DQ2-sec	PQQSFPQQP	
DQ8-glia-α1	QGSFPQPSQ	9mer match allowing three mismatches excluding Q/E at first and last position
DQ8-glia-γ1a	QQPQQPFPQ	
DQ8-glia-γ1b	QQPQQPYQP	9mer match allowing three mismatches
DQ8-glut-H1	QGYPTSPQ	

guideline, only 21 HLA-DQ2 restricted epitopes were used as patterns to scan all the protein sequences for exact 9mer matches using the FUZZPRO program in Emboss Package V6.4.0 (<http://emboss.sourceforge.net/>). The remaining one plus the four HLA-DQ2 restricted epitopes from ProPepper were used as patterns to scan all the protein sequences for 9mer matches but allowing three mismatches as recommended by the EFSA guideline. The four HLA-DQ8 restricted epitopes were also used as patterns to scan all the protein sequences for 9mer matches allowing three mismatches but excluding “E” (glutamic acid) or “Q” (glutamine) at the first and last position of a 9mer match as recommended by the EFSA guideline.

2.3. Search for the presence of HLA-DQ2 motif and peptide analysis

The Q/E-X1-P-X2 motif present in the HLA-DQ2 restricted epitopes was used as a pattern to identify its presence in all the protein sequences using the FUZZPRO program. Since a HLA-DQ2 restricted epitope is typically nine amino acids long with the motif generally starting at position 4 or 6, all of the 9mers containing motifs starting at position 4 or 6 were extracted from all the protein sequences containing the motif, followed by scanning for the presence of a proline duplex (PP), a null motif “ELPF”, positively charged amino acids (lysine, histidine, arginine) at positions 1, 4, 6, 7, and/or 9, glycine (G) or alanine (A) before the motif, serine after the motif (only applicable to the 9mers with motif starting at position 4), and a trypsin site (Naegeli et al., 2017). The presence of these amino acids or a trypsin site dramatically reduces the risk that the 9mer is a potential restricted epitope for celiac disease.

3. Results

When the HLA-DQ2 restricted epitopes were used to search for exact matches in the proteins from food plants and animals, there were no

Table 2

Sequence identity search results using celiac disease HLA-DQ2/DQ8 restricted epitopes, Q/E -X1-P-X2 motif, and protein sequences derived from the whole genome sequences of food plants and animals.

Organism	Total number of proteins analyzed	HLA-DQ2 9mer exact match				HLA-DQ2/DQ8 9mer match allowing three mismatches excluding Q/E at the first and last position				HLA-DQ2 Q/E-X1-P-X2 motif match			
		Number of proteins with ≥ one match	%	Total matches	Average matches per protein	Number of proteins with ≥ one match	%	Total Matches	Average matches per protein	Number of proteins with ≥ one match	%	Total Matches	Average matches per protein
Triticum Urartu	24 169	9	0.04	11	1.22	345	1.43	1257	3.64	2825	11.69	3388	1.20
Maize	58287	0	N/A ^a	N/A	N/A	1101	1.89	1921	1.74	8204	14.08	9666	1.18
Rice	41070	0	N/A	N/A	N/A	835	2.03	1513	1.81	5897	14.36	6982	1.18
Soybean	71523	0	N/A	N/A	N/A	1536	2.15	3323	2.16	9430	13.18	10926	1.16
Potato	37960	0	N/A	N/A	N/A	657	1.73	1129	1.72	5207	13.72	6604	1.27
Sunflower	73839	0	N/A	N/A	N/A	1570	2.13	3231	2.06	9373	12.69	10789	1.15
Tomato	36008	0	N/A	N/A	N/A	640	1.78	1132	1.77	5012	13.92	6132	1.22
Apple	60544	0	N/A	N/A	N/A	1051	1.74	2020	1.92	7730	12.77	8986	1.16
Banana	47707	0	N/A	N/A	N/A	997	2.09	2039	2.05	7003	14.68	8282	1.18
Cattle	29199	0	N/A	N/A	N/A	1105	3.78	2415	2.19	6954	23.82	9986	1.44
Swine	63575	0	N/A	N/A	N/A	2976	4.68	7287	2.45	16753	26.35	24786	1.48

^a Not applicable.

Table 3

Sequence identity search results using celiac disease HLA-DQ2/DQ8 restricted epitopes, Q/E-X1-P-X2 motif, and protein sequences of food plants and animals downloaded from UniProtKB database.

Organism	Total number of proteins analyzed	HLA-DQ2 9mer exact match				HLA-DQ2/DQ8 9mer match allowing three mismatches excluding Q/E at the first and last position				HLA-DQ2 Q/E-X1-P-X2 motif match			
		Number of protein with ≥ one match	%	Total matches	Average matches per protein	Number of protein with ≥ one match	%	Total matches	Average matches per protein	Number of protein with ≥ one match	%	Total matches	Average matches per protein
Barley	2303 (0) ^a	12	0.52	14	1.17	66	2.87	1222	18.52	222	9.64	531	2.39
Rye	702 (28; 3.99%)	38	5.41	54	1.42	64	9.12	4342	67.84	100	14.25	1012	10.12
Wheat	145925 (1288; 0.88%)	964	0.66	1419	1.47	4440	3.04	69072	15.56	22550	15.45	38563	1.71
Oats	318 (21; 6.60%)	19	5.97	19	1.00	52	16.35	365	7.02	58	18.24	127	2.19
Maize	132381	0	0.00	N/A ^b	N/A	2507	1.89	4642	1.85	18250	13.79	21730	1.19
Rice	168714	0	0.00	N/A	N/A	2592	1.54	4924	1.90	18710	11.09	22037	1.18
Potato	55642	0	0.00	N/A	N/A	597	1.07	951	1.59	4983	8.96	5624	1.13
Peanut	1462	0	0.00	N/A	N/A	20	1.37	23	1.15	115	7.87	128	1.11
Sunflower	54096	0	0.00	N/A	N/A	664	1.23	1239	1.87	5195	9.60	5903	1.14
Apple	2306	0	0.00	N/A	N/A	65	2.82	110	1.69	208	9.02	240	1.15
Banana	36637	0	0.00	N/A	N/A	435	1.19	811	1.86	3770	10.29	4280	1.14
Chicken	36239	0	0.00	N/A	N/A	854	2.36	1741	2.04	6706	18.50	8687	1.30
Cattle	32997	0	0.00	N/A	N/A	781	2.37	1667	2.13	5809	17.60	7649	1.32

^a Numbers inside the () are the protein sequences with gliadin or glutenin in the sequence definition field and its percentage of the analyzed protein sequences.

^b Not applicable.

matches except for proteins from the four known food sources known to affect celiac disease patients, and red wild einkorn wheat *urartu* which is a progenitor of bread wheat (Table 2 and Table 3). However, the searches for the five HLA-DQ2 (two without the Q/E-X1-P-X2) and DQ8 epitope matches allowing 3-mis matches indicated that more than 1–5% of the proteins from the *in silico* translation of coding sequences in the whole genome and UniProtKB database contain at least one such match with an average of close to two matches per protein. As expected, the proteins in UniProtKB database from the four food plants known to induce celiac disease showed a higher percentage of HLA-DQ2/DQ8 matches even though the number of proteins associated with gliadin or glutenin from these four sources is not high (Tables 2 and 3).

When searching for the presence of the Q/E-X1-P-X2 motif, the scenario changes dramatically. More than 12% of the proteins from the *in silico* translation of the coding sequences in the whole genomes of

non-celiac associated organisms contain the motif, ranging from 12.69% (sunflower) to 26.35% (swine) (Table 2). For the proteins in UniProtKB database, the number of proteins containing the motif ranged from 9.46% (barley) to 18.24% (oats) in the four food plants known to induce celiac disease, and from 7.87% (peanuts) to 18.50% (chicken) for the other food plants and animals (Table 3). This result is not surprising because the lower the number of amino acids in a match pattern, the higher the chances of finding a random match in a given protein sequence.

According to published literature (Tye-Din et al., 2010; Sollid et al., 2012) and EFSA guidance, the Q/E-X1-P-X2 motif generally starts at position 4 or 6 in a nine-amino acid-long HLA-DQ2 restricted epitope and the presence of certain amino acids such as positively charged amino acid residues make it unlikely for a 9mer peptide to bind to the HLA-DQ2 receptor. To understand how many 9mers in the proteins

Table 4

Analysis of all the 9mers containing HLA-DQ2 Q/E-X1-P-X2 motif in the protein sequences from the *in silico* translation of coding sequences in the whole genome sequences of food plants and animals.

Organism	Total number of 9mers containing Q/E-X1-P-X2 motif starting at position 4 or 6 ^a	Number of 9mers possibly subject to peptide modeling ^b	%
<i>Triticum urartu</i>	6682	3862	57.80
Maize	19080	10297	53.97
Rice	13748	7736	56.27
Soybean	21563	12456	57.77
Sunflower	21244	11766	55.39
Tomato	12120	6853	56.54
Apple	17775	10115	56.91
Potato	13060	7493	57.37
Banana	16347	9084	55.57
Cattle	19758	11272	57.05
Swine	49178	27485	55.89

^a The number of 9mers was derived from the total number of proteins containing the Q/E-X1-P-X2 motif in a 9mer format with the motif starting at position 4 or 6. In this case, there were possibly two relevant 9mers for each motif presence, but a 9mer with the motif starting at the first, second or third amino acid in a given protein was excluded; a 9mer with a motif being the last four amino acids of a given protein only has one relevant 9mer.

^b The number of 9mers was derived by subtracting the number of 9mers containing a proline duplex (PP), a null motif “ELPF”, positively charged amino acids (lysine, histidine, arginine) at positions 1, 4, 6, 7, and/or 9, glycine (G) or alanine (A) before the motif, serine after the motif (only applicable to the 9mers with motif starting at position 4), or a trypsin site from the total number of 9mers containing Q/E-X1-P-X2 motif starting at position 4 or 6.

Table 5

Analysis of all 9mers containing HLA-DQ2 Q/E-X1-P-X2 motif in the protein sequences of food plants and animals downloaded from the UniProtKB database.

Organism	Total number of 9mers containing Q/E-X1-P-X2 starting at position 4 or 6 ^[1]	Number of 9mers possibly subject to peptide modeling ^[2]	%
Barley	1037	809	78.01
Rye	2021	1895	93.77
Wheat	76465	52560	68.74
Oat	253	229	90.51
Maize	42945	24291	56.56
Rice	43231	24392	56.42
Potato	11058	6214	56.19
Peanut	253	139	54.94
Sunflower	11629	6651	57.19
Apple	471	280	59.45
Banana	8423	4680	55.56
Chicken	17195	9315	54.17
Cattle	15108	8622	57.07

^{[1][2]} As noted under Table 4.

containing the motif could be classified as potential restricted epitopes that are subject to the modeling step of a HLA-DQ2 peptide according to the EFSA guideline, all the 9mers containing the motif at position 4 or 6 were extracted and evaluated for the presence of a set of parameters, e.g. specific amino acid residues that will deteriorate the binding of a 9mer to HLA-DQ2 receptors. The results indicated that > 50% of the 9mers containing the motif at position 4 or 6 in the food plants and animals could be subject to HLA-DQ2 peptide modeling for potential binding to HLA-DQ2 receptors, no matter whether they were from the *in silico* translation of the coding sequences in the whole genome or UniProtKB database (Table 4 and Table 5), despite the fact that these food plants and animals are well known not to induce celiac disease. As expected, the percentage of such 9mers from the four food plants known to cause celiac disease was high at > 68%. This observation clearly indicates that the probability of being required to model a HLA-

DQ2 peptide based on a Q/E-X1-P-X2-motif-containing 9mer is very high. If the 9mer match criteria for HLA-DQ2 and 9mer match allowing three mismatches for HLA-DQ8 are implemented as a regulatory requirement to evaluate proteins for potential celiac disease risk, searches for the presence of the Q/E-X1-P-X2-motif will likely generate a lot of biologically meaningless random (false positive) hits with very low value in the risk assessment for potential celiac disease risk. This is a known issue with searches for matches of short peptides as exemplified by the abandonment of searches for 6 amino acid matches to known IgE allergen sequences as a method to evaluate potential cross-reactive allergenic risk (Stadler and Stadler, 2003; Silvanovich et al., 2006; Goodman et al., 2008; Herman et al., 2009; EFSA, 2010, 2011).

4. Discussion

Celiac disease symptoms include malabsorption of nutrients in the intestine, failure to thrive, diarrhea, and stomach pain due to the flattening of the intestinal mucosa. It is well known that celiac disease is caused by the gluten proteins in wheat, gluten-like hordeins in barley, and secalins in rye (Kagnoff, 2007; Peter, 2007; Scanlon and Murray, 2011; Fasano and Catassi, 2012). Several studies also demonstrated that oats are occasionally unsafe to celiac disease patients (Garsed and Scott, 2007). In the past two decades, a weight-of-evidence approach has been developed for evaluation of newly expressed proteins in GM crops for potential allergenicity. Evaluation of the newly expressed proteins in GM crops for potential allergenicity includes, but is not limited to, bioinformatics analysis of protein and reading frame sequences, simulated gastric fluid (SGF) and simulated intestine fluid (SIF) digestion using substrate proteins produced by microbial system equivalent to the *in planta* counterpart or derived from food source such as grain (James D. Astwood et al., 1996; Fu et al., 2002; Thomas et al., 2004; Ladics et al., 2011; Song, 2014; Schafer et al., 2016). In the bioinformatics analysis of newly expressed proteins for potential allergenicity, a peer-reviewed database including known allergens causing allergic reactions elicited by IgE and non-IgE mechanisms is generally searched for sequence similarity. Most non-IgE allergens are glutenins and gliadins from wheat, barley, and rye. For example, the widely used allergen COMPARE database (<http://comparedatabase.org/database/>) and FARRP/AllergenOnline database (Food Allergy Resource and Research Program at University of Nebraska, Lincoln; <http://www.allergenonline.org/>) (Goodman et al., 2016) contain 112/113 glutenins/gliadins sequences from wheat, barley, and rye. The majority of these sequences are gliadin, hordein, and secalin that are associated with celiac disease. Thus, evaluation of newly expressed proteins for potential celiac disease risk is already included in the current bioinformatics analysis for protein allergenicity assessment using a very conservative criterion: > 35% sequence identity over ≥ 80aa window. In addition, bioinformatics tools for celiac disease risk assessment have been developed by FARRP/AllergenOnline (<http://www.allergenonline.org/celiachome.shtml>) (Goodman et al., 2013). The tools include two parts: (1) identification of exact 9mer matches between a query protein and 1013 peptides associated with celiac disease, but with a precaution that exact identity matches to a 9mer peptide need to be interpreted carefully as they will likely over-estimate peptides that pose a risk of celiac disease risk; (2) a FASTA search of a query sequence against the 72 celiac disease associated proteins that are the sources of the 1013 peptides.

A search for the presence of the Q/E-X1-P-X2 in proteins from plant and animals used in food and not known to cause celiac disease resulted in hit rates ranging from 7.87% (proteins in UniProtKB database from peanuts) to 26.35% (from the *in silico* translation of the coding sequences in the whole genomes of swine). This means that it can be envisaged that approximately 10–25% of safe newly expressed proteins in GM crops could require further investigation in cases where no history of safe use can be shown. Even after the application of the recommended additional filters, EFSA guidance would require more than

half of the 9mers with a Q/E-X1-P-X2 motif to be modeled for potential binding to HLA-DQ2 receptors. The 2010 EFSA allergenicity guidance document (EFSA, 2010) indicated that accuracy, sensitivity, specificity, and feasibility of new bioinformatics tools should be evaluated before being routinely used for risk assessment. We conclude that implementation of a regulatory requirement resulting in a high false-positive rate does not help to increase public trust in the risk assessment, but might rather result in unnecessary uncertainty by the public regarding the safety of GM crops.

As current scientific evidence limits celiac disease incidence to the intake of grains from wheat, barley, rye, oats, and their closely related species such the wheat subfamily (*Pooideae*), it is recommended to focus on genes derived from these known food sources in the evaluation of newly expressed proteins in GM food crops for potential risk to elicit celiac disease. Furthermore, if further analysis is needed, we suggest the use of the identification of a 9mer exact match between a query protein and the known celiac disease induction peptides as recommended by the EFSA guideline along with a supplementary sequence comparison as suggested by FARRP/AllergenOnline to optimize sensitivity versus specificity in celiac disease risk assessment of newly expressed proteins in GM crops.

Transparency document

Transparency document related to this article can be found online at <https://doi.org/10.1016/j.yrtph.2018.09.021>

References

- Astwood, J.D., Leach, J.N., Fuchs, R.L., 1996. Stability of food allergens to digestion *in vitro*. *Nat. Biotechnol.* 14, 1269–1273.
- Codex Alimentarius Commission, 2009. Foods derived from modern biotechnology. www.fao.org/docrep/011/a1554e/a1554e00.htm.
- EFSA, 2010. Scientific opinion on the assessment of allergenicity of GM plants and microorganisms and derived food and feed. *EFSA J.* 8, 1700.
- EFSA, 2011. Guidance for risk assessment of food and feed from genetically modified plants. *EFSA J.* 9, 2150.
- Food and Agriculture Organization of the United Nations, 2001. Evaluation of Allergenicity of Genetically Modified Foods. Report of Joint FAO/WHO Expert Consultation. Rome.
- Fasano, A., Catassi, C., 2012. Clinical practice. Celiac disease. *N. Engl. J. Med.* 367, 2419–2426.
- Fu, T.T., Abbott, U.R., Hatzos, C., 2002. Digestibility of food allergens and nonallergenic proteins in simulated gastric fluid and simulated intestinal fluid – a comparative study. *J. Agric. Food Chem.* 50, 7154–7160.
- Garsed, K., Scott, B.B., 2007. Can oats be taken in a gluten-free diet? A systematic review. *Scand. J. Gastroenterol.* 42, 171–178.
- Goodman, R., Amnuaycheewa, P., Wise, J., Taylor, S.L., 2013. Food allergy and anaphylaxis – 2039. Bioinformatics evaluation of new proteins in genetically engineered organisms and novel foods for potential risks of food allergy and celiac elicitation. *World Allergy Organ. J.* 6, 408.
- 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., 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, 73–81.
- Herman, R.A., Song, P., Thirumalaiswamysekhar, A., 2009. Value of eight-amino-acid matches in predicting the allergenicity status of proteins: an empirical bioinformatic investigation. *Clin. Mol. Allergy* 7, 9.
- ISAAA, 2017. Global Status of Commercialized Biotech/GM Crops in 2017: Biotech Crop Adoption Surges as Economic Benefits Accumulate in 22 Years. ISAAA Brief No. 53. ISAAA, Ithaca, NY.
- Kagnoff, M.F., 2007. Celiac disease: pathogenesis of a model immunogenetic disease. *J. Clin. Invest.* 117, 41–49.
- Kathage, J., Qaim, M., 2012. Economic impacts and impact dynamics of bt (*Bacillus thuringiensis*) cotton in India. *Proc. Natl. Acad. Sci. Unit. States Am.* 109, 11652–11656.
- Klümper, W., Qaim, M., 2014. A meta-analysis of the impacts of genetically modified crops. *PLoS One* 9, e111629.
- 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.* 60, 46–53.
- Naegeli, H., Birch, A.N., Casacuberta, J., De Schrijver, A., Gralak, Mikolaj A., Guerche, P., Jones, H., Manachini, B., Messéan, A., Nielsen, E.E., Nogué, F., Robaglia, C., Rostoks, N., Sweet, J., Tebbe, C., Visioli, F., Wal, J.M., Eigenmann, P., Epstein, M., Hoffmann-Sommergruber, K., Koning, F., Lovik, M., Mills, C., Moreno, F.J., van Loveren, H., Selb, R., Fernandez Dumont, A., 2017. Guidance on allergenicity assessment of genetically modified plants. *EFSA J.* 15.
- National Research Council, 2010. The Impact of Genetically Engineered Crops on Farm Sustainability in the United States. Academy Press, Washington, DC.
- Peter, H.R., Green, 2007. Celiac disease. *N. Engl. J. Med.* 357, 1731–1743.
- Scanlon, S.A., Murray, J.A., 2011. Update on celiac disease – etiology, differential diagnosis, drug targets, and management advances. *Clin. Exp. Gastroenterol.* 4, 297–311.
- Schafer, B.W., Embrey, S.K., Herman, R.A., 2016. Rapid simulated gastric fluid digestion of in-seed/grain proteins expressed in genetically engineered crops. *Regul. Toxicol. Pharmacol.* 81, 106–112.
- Silvanovich, A., Nemeth, M.A., Song, P., Herman, R., Tagliani, L., Bannon, G.A., 2006. The value of short amino acid sequence matches for prediction of protein allergenicity. *Toxicol. Sci.* 90, 252–258.
- Sollid, L.M., Qiao, S.W., Anderson, R.P., Gianfrani, C., Koning, F., 2012. Nomenclature and listing of celiac disease relevant gluten t-cell epitopes restricted by HLA-DQ molecules. *Immunogenetics* 64, 455–460.
- Song, P., 2014. Bioinformatics applications in regulatory assessment for potential allergenicity of transgenic proteins in food. In: Watson, Ronald Ross, Preedy, Victor R. (Eds.), *Genetically Modified Organisms in Food - Production, Safety, Regulation and Public Health*, Ch.35. Academic Press.
- Stadler, M.B., Stadler, B.M., 2003. Allergenicity prediction by protein sequence. *Faseb. J.* 17, 1141–1143.
- 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., Holsapple, 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.* 39, 87–98.
- Tye-Din, J.A., Stewart, J.A., Dromey, J.A., Beissbarth, T., van Heel, D.A., Tatham, A., Henderson, K., Mannering, S.I., Gianfrani, C., Jewell, D.P., Hill, A.V., McCluskey, J., Rossjohn, J., Anderson, R.P., 2010. Comprehensive, quantitative mapping of T cell epitopes in gluten in celiac disease. *Sci. Transl. Med.* 2, 41ra51.
- Wieser, H., 2007. Chemistry of gluten proteins. *Food Microbiol.* 24, 115–119.