Food and Chemical Toxicology 50 (2012) 3741–3751

Contents lists available at SciVerse ScienceDirect

Food and Chemical Toxicology

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

Assessment of possible allergenicity of hypothetical ORFs in common food crops using current bioinformatic guidelines and its implications for the safety assessment of GM crops

Gregory J. Young *, Shiping Zhang, Henry P. Mirsky, Robert F. Cressman, Bin Cong, Gregory S. Ladics, Cathy X. Zhong

Pioneer Hi-Bred International, Inc., DuPont Agricultural Biotechnology, Wilmington, DE 19880, USA

ARTICLE INFO

Article history:
Received 18 April 2012
Accepted 20 July 2012
Available online 31 July 2012

Keywords:
Genetically modified crops
Bioinformatics
Codex Alimentarius Commission
Regulatory recommendations
Assessment of allergenicity
Food safety

ABSTRACT

Before a genetically modified (GM) crop can be commercialized it must pass through a rigorous regulatory process to verify that it is safe for human and animal consumption, and to the environment. One particular area of focus is the potential introduction of a known or cross-reactive allergen not previously present within the crop. The assessment of possible allergenicity uses the guidelines outlined by the Food and Agriculture Organization (FAO) and World Health Organization's (WHO) Codex Alimentarius Commission (Codex) to evaluate all newly expressed proteins. Some regulatory authorities have broadened the scope of the assessment to include all DNA reading frames between stop codons across the insert and spanning the insert/genomic DNA junctions. To investigate the utility of this bioinformatic assessment, all naturally occurring stop-to-stop frames in the non-transgenic genomes of maize, rice, and soybean, as well as the human genome, were compared against the AllergenOnline (www.allergenonline.org) database using the Codex criteria. We discovered thousands of frames that exceeded the Codex defined threshold for potential cross-reactivity suggesting that evaluating hypothetical ORFs (stop-to-stop frames) has questionable value for making decisions on the safety of GM crops.

1. Introduction

Genetically modified (GM) crops have been safely grown and consumed across most regions in the world for the past 16 years

tion surpasses 7 billion people, and water scarcity, soil salinity, and other abiotic and biotic stresses continue to increase, the need to produce and maintain high yields for a variety of important food crops through techniques such as genetic engineering will become even more of an imperative. However, before a GM crop can be commercialized it must pass through a rigorous regulatory process to verify that it is safe for human and animal consumption, and to the environment. One particular area of concern is for the potential introduction of a known or cross-reactive allergen not previously present within the crop. In order to verify that a newly expressed protein is unlikely to be an allergen or will not have IgE cross-reactivity with any allergens, regulatory authorities require a thorough bioinformatic assessment of potential allergenicity following guidelines outlined by the Food and Agriculture Organization and World Health Organization's Codex Alimentarius Commission (2009), herein referred to as Codex. Assuming the newly expressed protein is not from a known allergenic source, the initial assessment involves an *in silico* search for amino acid sequence homology to a database of known allergens using either the FASTA or BLAST algorithms (Pearson and Lipman, 1988; Stephen et al., 1997). Cross-reactivity is considered a possibility if more than 35% identity over a length of 80 or more amino acids (35%/⩾80aa) is shared between the protein and the allergen. In addition, the guidelines also specify performing a stepwise contiguous identical amino acid search using a "scientifically justifiable" length in order to identify any short linear epitopes that may be present in the newly expressed protein sequence (Codex, 2009). The standard practice for the industry is to use a length of eight or more residues when searching for contiguous identical amino acid matches based on past experiments and the current requirements from many different regulatory agencies around the world (Hileman et al., 2002; Stadler and Stadler, 2003; Ladics et al., 2006; Silvanovich et al.,

Abbreviations: aa, amino acid; GM, genetically modified; FAO, Food and Agriculture Organization of the United Nations; WHO, World Health Organization of the United Nations; FARRP, Food Allergy Research and Resource Program; EFSA, European Food Safety Authority; ORF, open reading frame; IgE, immunoglobulin E.

* Corresponding author. Address: DuPont Agricultural Biotechnology, Pioneer Hi-Bred International, Experimental Station Bldg. E353, Rt. 141 & Henry Clay Wilmington, DE 19880, USA. Tel.: +1 302 695 6954.

E-mail address: Gregory.Young@cgr.dupont.com (G.J. Young).

0278-6915
http://dx.doi.org/10.1016/j.fct.2012.07.044

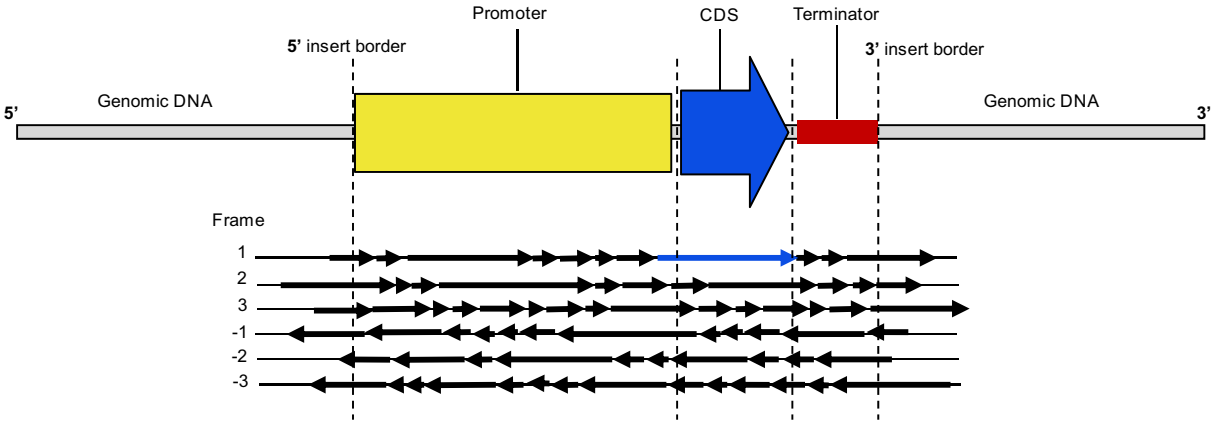

Fig. 1. Schematic representation of a single gene insertion containing a promoter (yellow), a coding sequence (CDS; blue), and a terminator sequence (red). Flanking genomic DNA is shown as the grey horizontal bar on the 5′ and 3′ ends. All stop-to-stop frames are shown below the insertion in all six DNA reading frames. The stop-to-stop frames are shown as arrows pointing in the 5′–3′ direction. The blue arrow represents the stop-to-stop frame of the intended coding sequence of the transgene. The insert/genomic junctions and genetic element junctions are designated by vertical dashed lines. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

2006; Ladics et al., 2011). If a protein exceeds the 35%/⩾80aa threshold or has a contiguous identical match of eight residues or more, then additional tests up to and including screening for IgE binding using allergic patient sera are required to evaluate the cross-reactivity potential (Codex, 2009).

In addition to applying the Codex guidelines to the newly expressed proteins in a GM crop, some authorities have broadened the scope of the allergenicity assessment to include all sequences between stop codons from all six DNA reading frames present within the insert and spanning the insert/genomic DNA junction sites with no minimum length (EFSA, 2011). Fig. 1 shows a schematic representation of what such an analysis would include for a simple, single-gene insertion containing a promoter, a coding sequence (CDS), and a terminator sequence. Under these requirements all stop-to-stop reading frames shown in Fig. 1 within the insert or spanning the insert/genomic junction are evaluated for sequence homology to known allergens. Any requirements for potential gene expression of the hypothetical ORF (defined as sequence between two stop codons), such as the presence of a translational start codon, promoter, splice sites, and/or transcriptional terminator sequence are not a prerequisite for conducting the bioinformatic analysis. Rather the concern is with the theoretical potential that such hypothetical ORFs may be cross-reactive or act as allergens if it were ever to be placed in a context in which it would be expressed.

The objective of this study was to investigate the utility of evaluating hypothetical ORFs as it relates to making decisions on the safety of a GM plant. To accomplish this we evaluated the allergenicity potential of all stop-to-stop frames within the reference genomes of non-transgenic soybean, maize, and rice using the criteria defined in the Codex guidelines (2009). We also evaluated the human genome to serve as a non-plant control with a large genome and no known allergens. For rice, two sub-species varieties, for which complete genome sequence was available, were analyzed and compared in order to demonstrate the natural variation in hypothetical ORF content that could be expected between crop varieties without differences in novel allergen content. Several loci were also compared between non-transgenic maize inbred lines B73 and Mo17 for the same purpose. Because stop-to-stop frames represent hypothetical peptide or polypeptide sequences and are not true open reading frames (ORFs) (i.e. they do not contain the appropriate translational start signal), they are hereafter referred to as novel polypeptides in order to avoid any confusion with expressed or predicted protein-coding sequence even though at times the novel polypeptide sequence may overlap with a expressed or protein-coding sequence. All novel polypeptides translated from the genome datasets greater than or equal to 80aa in length (⩾80aa) were analyzed for sequence identity to the AllergenOnline database of allergens version 11 (www.allergenonline.org) (also referred to as the FARRP 11 database) using the FASTA algorithm (Pearson and Lipman, 1988). In addition, all eight residue identical matches to the FARRP11 database were identified from the set of novel polypeptides equal to or greater than eight amino acids in length from all five genomes analyzed in order to assess the frequency of such matches within these unregulated common food crops. The results show that thousands of endogenous novel polypeptides exceeded the 35%/⩾80aa threshold for potential cross-reactivity, that novel polypeptides can be highly variable between different varieties and breeding lines, and that eight residue identical matches are highly common in each of the genomes analyzed. The observed results are not consistent with the safe history of these common food crops, the small number of known endogenous allergens, and what is general known about the different families of plant food allergens (Breiteneder and Radauer, 2004). As a result, the value of analyzing all stop-to-stop frames within and spanning the insertion site as part of the safety assessment of a GM crop, along with the search for short continuous identical matches of eight residues, is questioned.

2. Materials and methods

The translation of stop-to-stop frames, the alignments to the AllergenOnline database using FASTA, and the search for eight residue identical matches were performed in a manner consistent with the requirements from regulatory authorities for the bioinformatics safety assessment of the transgene insertion and the region spanning the 5′ and 3′ insert/genomic junctions.

2.1. Databases

The publically available genome sequences, along with an annotated set of protein and coding sequence of three common food crops (maize, soybean, and rice) were used for this study. The *Zea mays* L. B73 (RefGen_2, 5a.59 annotation release), the *Glycine max* Williams82 (version Glyma1.0), and *Oryza sativa* subspecies *japonica* MSU release 6.0 were downloaded from the Phytozome website (Schnable et al., 2009; Schmutz et al., 2010; Ouyang et al., 2007; IRGSP, 2005; www.phytozome.org). A second rice genome was also analyzed. The *Oryza sativa* subspecies *indica* assembly was downloaded from the Beijing Genomics Institute (rice.genomics.org.cn; Yu et al., 2002). The human genome dataset and protein sequences were downloaded from National Center for Biotechnology Information (NCBI) (ftp://ftp.ncbi.nlm.nih.gov/genomes/H_sapiens/). The protein set only included the reference sequences (RefSeq) that was exported as part of the genome

annotation process. For the genome sequence, only the reference assembly GRCH37.p5 and the assembled chromosomes were utilized. The un-localized and alternative loci were not included in our analysis. All datasets were unfiltered and not masked for repetitive content because it is not allowed by the regulations. The number of nucleotides that were used in our analysis is shown in the first column of Table 1. In addition, the UniGene set of expressed sequences for each species (*Z. mays*, *G. max*, and *O. sativa*) was downloaded from NCBI (ftp://ftp.ncbi.nih.gov/repository/Unigene/). The set of allergens was from the AllergenOnline database version 11 (http://www.allergenonline.org/). The AllergenOnline database was developed and maintained by the Food Allergy Research and Resource Program (FARRP) at the University of Nebraska in Lincoln. It includes a comprehensive peer-reviewed list of allergens and is commonly used when performing allergen assessments for regulatory submissions. The version 11 database contained a total of 1491 sequences and is hereafter referred to as the FARRP 11 database.

2.2. Stop-to-stop translations in all frames and the FASTA35 analysis

The novel polypeptide sequences were created by scanning all six DNA reading frames (starting at base 1, 2 or 3 of each DNA sequence and translating each non-overlapping nucleotide triplet into an amino acid using the standard genetic codon table, then reversing the sequence and repeating the process) from each chromosome/scaffold or contig and saving all sequences that are greater than or equal to 8aa between stop codons. This is the same method for which stop-to-stop frames would be translated for the evaluation of a transgenic insertion (Fig. 1), but on a much larger scale. Because the Codex guidelines require a minimum alignment length of 80aa, only the fraction of novel polypeptides greater than or equal to 80aa in length were analyzed for sequence homology to the FARRP 11 allergen database using FASTA_v3.5 (FASTA35) (Pearson and Lipman, 1988). FASTA was chosen rather than BLAST, because it is a more sensitive algorithm and is used for the bioinformatic safety assessment reports sent to regulatory agencies. The default FASTA settings were used except for the histogram feature was inactivated. Gaps were considered as mismatches in the calculation of percent identity as is the default for the FASTA and therefore, a few novel polypeptides less than 80 residues in length, which may have generated an alignment of 80 or more with the gaps included, would have been missed by the analysis. A sliding window of 80aa was not used because it is known to artificially increase the false positive rate and potentially miss significant matches by limiting the alignment length (Ladics et al., 2007; Cressman and Ladics 2009). All the alignments generated by FASTA were scanned for 35.0% or greater amino acid identity with an alignment length of 80 residues or greater (35%/⩾80aa). An example of a FASTA generated alignment from this study is shown in Fig. 2. The number of novel polypeptides with one or more alignments exceeding the 35%/⩾80aa threshold was counted and the top scoring alignment parsed into a table for each genome dataset. From this table the novel polypeptide length, *E*-value distributions, and allergen frequency were gathered for the top-scoring alignment.

The number of novel polypeptides that overlapped with a protein-coding sequence frame was estimated by comparing the set of novel polypeptides that had one or more alignments which exceeded the 35%/⩾ 80aa threshold to the publically available annotated set of proteins from each of the four plant genomes using BLASTp with an *E*-value cutoff at 0.001 (Stephen et al., 1997). The BLAST results were then parsed based on a cutoff of 95% identity or more across an alignment that was at least 80% or greater than the length of the novel polypeptide query. The fraction of alignments with lengths at 90%, 70%, 60%, 50%, and 10% of the query length were also parsed from the BLASTp results and are shown in Supplementary Table S1. The 80% query length cutoff provided a good estimate of the number of novel polypeptides derived from protein-coding sequence because and it allowed for some sequence to flank the exons (which would be necessary when looking at stop-to-stop frames), but also ensured that the majority of the alignment was from the protein-coding portion. In addition, in order to identify those novel polypeptides derived from an alternative frame of a protein-coding gene and/or expressed sequence, tBLASTn searches were performed using the publically available nucleotide coding sequence (CDS) and UniGene datasets at an *E*-value cutoff set at 1×10^{-5} (Stephen et al., 1997). The tBLASTn results were also parsed based on the 95% identity over at least 80% query length criteria. The novel polypeptides derived from alternative (non-coding) frames of protein-coding sequence were uniquely identified by cross-referencing the CDS tBLASTn matches with the protein-coding BLASTp matches and removing those that were redundant; whereas, the set of novel polypeptides unique to the UniGene dataset were identified by cross-referencing with both the CDS and protein-coding results (Table S2). The set of novel polypeptides with matches to a coding sequence or UniGene entry were added to the overall number of novel polypeptides derived from coding and expressed sequence (Table 1).

The number of unique protein-coding genes with a match to one or more novel polypeptides was determined by sorting the parsed BLASTp results (95% identity across at least 80% of the novel polypeptide sequence). The matches to the endogenous allergens were determined by searching each allergen from *Z. mays*, *G. max*, and *O. sativa* in the FARRP1 database against the parsed table of top-scoring alignments (Table 2). The same was done in order to determine the most frequent allergen matches seen in the top-scoring alignments across the entire FARRP11 database (Table 3).

2.3. Eight residue contiguous identical matches against the FARRP 11 database

The standard practice for evaluating short contiguous identical matches is to identify all possible unique eight amino acid (8-mer) segments (minimum length required by regulators) from the ORF in question and search for all identical matches to the allergen database. In this analysis, a set of eight amino acid sequences (i.e. 8-mer) was compiled by scanning each translated stop-to-stop frame equal to or greater than eight amino acids in length using a window length of eight, shifting one position at a time, and recording the sequence of the current window. Then number of times each unique 8-mer occurred within a genome was counted. Next, all allergens were screened against each unique 8-mer dataset using a custom made script that identified identical matches. In addition to the genome dataset, sets of 8-mer sequences were also complied and screened against the FARRP11 database from the annotated set of protein-coding sequences from each of the four plant genomes (Table 4). Using this method we have estimated the number of times any 8-mer from the AllergenOnline database appears in each of the five genomes and four protein-coding datasets.

2.4. Intraspecies comparison of novel polypeptide content

The set of greater than or equal to 80aa novel polypeptides that exceeded the 35%/⩾80aa threshold from the two rice subspecies *japonica* and *indica* were directly compared using a reciprocal BLASTp analysis. All novel polypeptides that matched at 95% identity or greater spanning an alignment length that was at least 80% of the length of the query sequence were considered as conserved sequences. Those that fell below 95% across at least 80% of the query length were considered as unique to one of the subspecies (Table 6). Under regulations, any difference between sequences (e.g. a single nucleotide change) is enough for the sequence to be considered as unique. Therefore, for the purposes of this analysis, setting a criterion of 95% identity over at least 80% of the query length was adequate. For the comparison between the maize lines B73 and Mo17, four previously characterized loci

Table 1
Summary of the novel polypeptide FASTA analysis against the FARRP 11 allergen database per genome.

Genome	Novel polypeptides ⩾80aa[a]	35%/⩾80aa[b]	35%/⩾80aa Match protein-coding gene[c]	35%/⩾80aa Match with CDS or UniGene[d]	35%/⩾80aa Total percent from coding sequence[e]	35%/⩾80aa Unique protein-coding genes[f]
Z. mays	6,993,325	33,475	1052	3041	12%	1188
G. max	1,434,209	3393	1397	474	55%	1282
O.s. ssp. indica	1,196,644	10,034	916	2115	32%	930
O.s. ssp. japonica	1,064,957	11,479	1089	3072	36%	1097
H. sapiens[g]	5,882,480	21,968	190	–	–	–

[a] The number of novel polypeptides that are ⩾80 amino acids in length translated from the reference genome sequence and included in the FASTA35 analysis.
[b] The number of unique ⩾80aa novel polypeptides that have one or more alignments that exceeded the 35%/⩾80aa threshold.
[c] The number of unique ⩾80aa novel polypeptides that that exceeded the 35%/⩾80aa threshold which matched a protein-coding sequence at ⩾95% identity across at least 80% of the length of the novel polypeptide sequence using BLASTp.
[d] The number of unique ⩾80aa novel polypeptides that exceeded the 35%/⩾80aa threshold which matched an non-coding frame of a coding sequence or matched a Unigene sequence at ⩾95% identity across at least 80% of the length of the novel polypeptide sequence using tBLASTn.
[e] The total percentage of ⩾80aa novel polypeptides that exceeded the 35%/⩾80aa threshold that overlapped with a coding or expressed sequence (columns 4 and 5 divided by column 3).
[f] The number of unique protein-coding genes with one or more matches to a ⩾ 80aa novel polypeptide that exceeded the 35%/⩾80aa threshold.
[g] The analysis using tBLASTn against nucleotide CDS and UniGene datasets was not performed for the human set of novel polypeptides.

```
 Query: 8-125682956-2.70340, 96 aa

>>gi|27806257|ref|NP_776945.1| collagen alpha-2(I) chain  (1364 aa)
 initn: 138 init1:  82 opt: 171  Z-score: 199.4  bits: 45.5 E(): 4e-06
Smith-Waterman score: 171; 48.2% identity (58.8% similar) in 85 aa
overlap (2-78:700-779)

                                                 10        20        30
8-1256                                        RAGPAG-RGFLEKRKKREGAVGPAGPKGGRG
                                               ::::: ::   .:    : :::::::.:  :
gi|278 DGARGAPGAIGAPGPAGANGDRGEAGPAGPAGPAGPRGSPGER----GEVGPAGPNGFAG
     670       680       690       700       710           720

               40          50          60        70         80
8-1256 ARLGSPGRPTA-GE-GWLGRRGAHAGRGG----GAAGPLGPRGRGGWA-KRGEGGREKEK
         :. :.: : :: :  : .: ..  :     :::::: :: :  : : .::.::
gi|278 PA-GAAGQPGAKGERGTKGPKGENGPVGPTGPVGAAGPSGPNGPPGPAGSRGDGGPPGAT
         730       740       750       760       770       780

            90
8-1256 AFLFLIFHIFLYA

gi|278 GFPGAAGRTGPPGPSGISGPPGPPGPAGKEGLRGPRGDQGPVGRSGETGASGPPGFVGEK
         790       800       810       820       830       840
```

Fig. 2. In this example of a FASTA generated alignment, the novel polypeptide 8-125682956-2.70340 from *Z. mays* was aligned to the collagen alpha-2(I) chain protein from the FARRP11 database over a length of 85 aa (shown in red, which does include gaps) with 48.2% identity. The two dots indicate an identity (41/85 aa overlap), whereas a single dot indicates a similar amino acid. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 2
The number allergens from each species in the FARRP database and the number of novel polypeptides with a species-specific allergen as the top scoring match exceeding the Codex 35%/⩾80aa threshold.

Genome	FARRP 11 database[a]	Top matching species-specific allergens[b]
Z. mays	24	123
G. max	36	171
O.s ssp. indica	20	110
O.s ssp. japonica	20	98

[a] The number of allergen accessions in the FARRP 11 database.
[b] The number of novel polypeptides for which a species specific allergen is the top scoring alignment in the FASTA analysis.

were used from Brunner et al. (2005). The names of these loci included are: 9002, 9008, 9009, and *bz1*. The Genbank accession numbers for these loci are as follows: AY664413, AY664414, AY664415, AY664416, AY664417, AY664418, AY664419, and AF448416. The stop-to-stop translation of the novel polypeptides was done using the EMBOSS program getorf (Rice et al., 2000), and the allergenicity assessment was done as previously described for the genome set of novel polypeptides using FASTA35. To determine if the novel polypeptides were conserved, the B73 and Mo17 sets were directly compared using a reciprocal BLASTp analysis in same manner as described above for rice. The retroelement annotations were based on the positions of the novel polypeptide within the annotated sequence provided by Brunner et al. (2005).

3. Results

3.1. The bioinformatic assessment of possible allergenicity of novel polypeptides using FASTA35 against the FARRP 11 database

All stop-to-stop frames equal to or greater than 24 nucleotides (or 8 amino acids) in length were translated from the publically available, unfiltered, genome datasets of *Z. mays* L., *G. max*, *O. sativa* L. ssp. *japonica*, *O. sativa* L. ssp. *indica*, and *Homo sapiens* (maize, soybean, rice sub-species japonica and indica, and human, respectively). The translated stop-to-stop frames are referred to as novel polypeptides in order to distinguish them from true protein-coding sequence. Because the Codex (2009) risk assessment for allergenicity threshold is set at a length of 80 amino acids, only those novel polypeptides equal to or greater than 80 amino acids were used for analysis against the FARRP 11 database. The genomes of *G. max* and *O.s.* (*Oryza sativa*) ssp. *japonica* and *indica* contained well over a million novel polypeptides 80 amino acids or more in length, while the corn and human genomes contained nearly seven and six million respectively (Table 1). This set of more than 16 million novel polypeptides was evaluated for sequence homology against the FARRP 11 database of allergens using FASTA35. An example of a FASTA alignment from the analysis is shown in Fig. 2.

Table 3
The three most frequent allergen matches from each 35%/⩾80aa novel polypeptide dataset.

Allergen	*Z. mays*	*O.s.* ssp. *indica*	*O.s.* ssp. *japonica*	*G. max*	*H. sapien*
Bos collagen alpha2[a]	13,532	2395	2931	–	11,737
Lol p 5[b]	3959	1959	2066	–	1206
Amb 4 defensin-like protein[c]	3624	1356	1887	–	1882
Cuc m 1[d]	–[g]	–	–	210	–
Putative leucine-rich repeat protein[e]	–	–	–	152	–
Asp f 9[f]	–	–	–	150	–

[a] Bos collagen alpha2 (*Bos Taurus*; Bovine; Vaccine).
[b] Lol p 5 (*Lolium perenne*; perennial ryegrass; Aero Plant).
[c] Amb 4 (*Ambrosia artemisiifolia*; short ragweed; Aero Plant).
[d] Cuc m 1(*Cucumis melo*; Muskmelon; Food Plant).
[e] Putative leucine-rich repeat protein (*Triticum aestivum*; wheat; Food Plant).
[f] Asp f 9 (*Aspergillus fumigatus*; Fungus; Aero Fungi)
[g] – Represents that this allergen is not in the top-three matches for that species.

Table 4
8-Mer matches to the FARRP 11 database.

Genome	8-Mers[a]	8-Mer matches[b]	Unique 8-mers novel polypeptides[c]	Unique 8-mers protein-coding[d]
Z. mays	662,345,694	217,874	17,615	6017
G. max	578,119,236	85,860	16,911	8961
O.s. ssp. indica	346,448,671	73,454	12,407	5816
O.s. ssp. japonica	326,736,944	71,332	12,056	5737
H. sapiens	2,180,452,518	234,419	31,938	–

[a] Total number of unique 8-mers within the set of novel polypeptides from each genome dataset.
[b] The frequency of 8-mers within the novel polypeptide set with identical matches to the FARRP 11 database.
[c] The number of unique 8-mers within the novel polypeptide set with one or more identical matches to the FARRP 11 database.
[d] The number of unique 8-mers within the set of protein-coding genes from each genome that has one or more matches to the FARRP 11 database. The set of protein-coding genes from humans was not analyzed.

Table 1 shows the number of novel polypeptides with one or more alignments that exceeded the Codex threshold of 35.0% identity across 80 amino acids or more (35%/⩾80aa) for each genome dataset. Maize, which contained the highest overall number of novel polypeptides, also had the highest number that exceeded the 35%/⩾80aa threshold at 33,475, while the genomes of the two rice varieties, *indica* and *japonica*, contained 10,034 and 11,479 above threshold matches, respectively. Interestingly, the *G. max* genome, twice as large and with a similar overall number of novel polypeptides as either of the two rice genomes, contained only a fraction of the number exceeding the threshold (3393). The presence of thousands of novel polypeptides exceeding the Codex threshold was not unique to plants—the human genome contained 21,968 novel polypeptides with above threshold matches to known allergens (Table 1). The size distribution of this set of novel polypeptides is shown in Fig. 3 for each genome. Greater than 90% of the novel polypeptides from each genome, with the exception of *G. max* at 85%, are between 80 and 300 amino acids in length. Soybean had a slightly higher fraction of novel polypeptides greater than 300 amino acids than the other four genomes.

In order to estimate the number of novel polypeptides (stop-to-stop frames) that overlapped with protein-coding sequence, the set of 35%/⩾80aa novel polypeptides shown in Table 1 was compared to the public annotated set of proteins using BLASTp for each of the four plant genomes. At the 80% query-length cutoff a total of 1052, 1397, 916, and 1089 novel polypeptides from *Z. mays*, *G. max*, *O.s.* ssp. *indica* and *japonica*, respectively, matched a protein-coding sequence (Table S1). This represents approximately 3.1%, 41.2%, 9.1%, and 9.5% of the total number of novel polypeptides that exceeded the Codex threshold from *Z. mays*, *G. max*, *O.s.* ssp. *indica* and *japonica*, respectively. A much smaller percentage of matches to protein-coding genes was observed from the human dataset, in which only 190 (0.8%) of the novel polypeptides had a match to one or more of the reference proteins. When the percent query length threshold was lowered, the number of novel polypeptides increased only slightly and appears to level out at approximately 60% (Table S1). When factoring in the set of 35%/⩾80aa novel polypeptides with matches to an alternative-frame of a coding sequence, or to a UniGene entry, to the set that overlapped with a protein sequence, the fraction of unique novel polypeptides that exceeded the Codex threshold derived from a frame spanning a coding and/or expressed sequence ranges from 12% in maize to 55% in soybean (Table 1). With the exception of soybean, the majority of the novel polypeptides that exceeded the 35%/⩾80aa threshold in maize and rice are from frames that do not span coding and/or expressed sequence. Though, in maize for example, the estimated 12% of novel polypeptides that did span a coding and/or expressed sequence still equals approximately four thousand frames that exceeded the Codex's safety threshold for potential allergenicity. Similar to the number of novel polypeptides that spanned one or more protein sequences, the number of unique protein-coding sequences that had a match to one or more novel

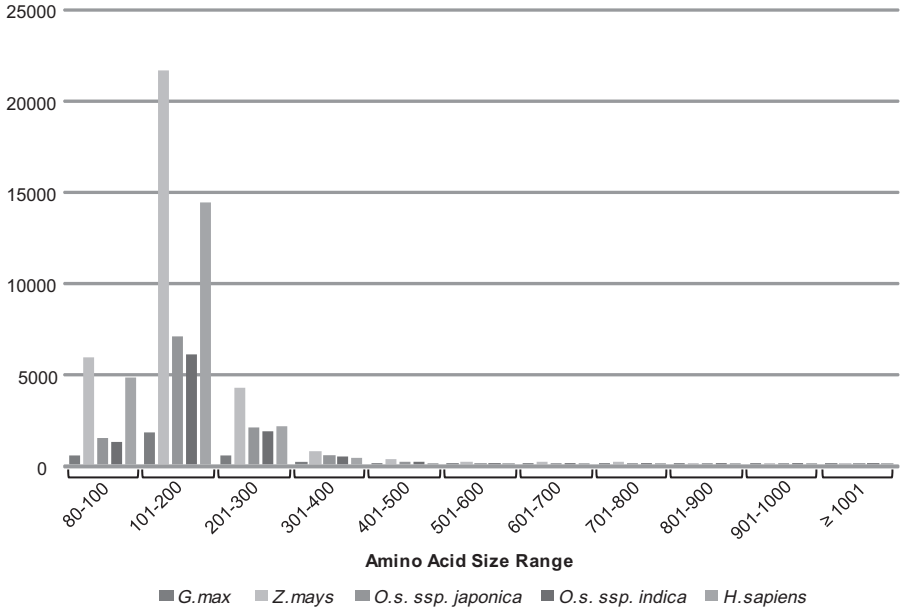

Fig. 3. Bar graph showing the amino acid size distribution of all novel polypeptides exceeding the Codex 35%/⩾80aa threshold.

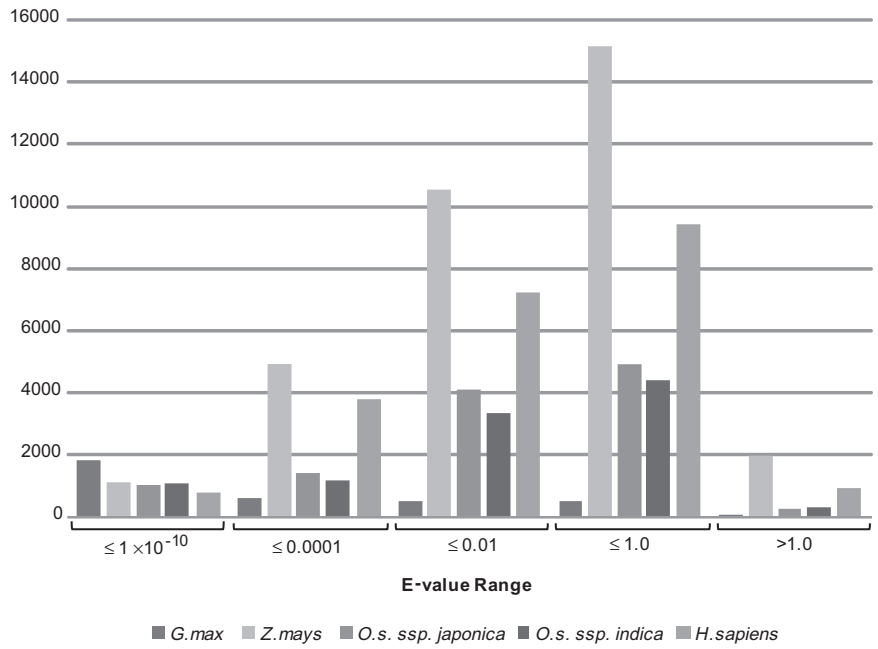

Fig. 4. Bar graph showing the number of novel polypeptides exceeding the Codex 35%/⩾80aa threshold per *E*-value of the top scoring alignment to the FARRP 11 database. The order of the bars from left to right: *G. max*, *Z. mays*, *O.s.* ssp. *japonica* and *indica*, and *H. sapiens*. *E*-value ranges are: $\leqslant 1 \times 10^{-10}$, between 1×10^{-10} and 0.0001, 0.0001 and 0.01, 0.01 and 1.0, and >1.0.

polypeptides above the Codex (35%/⩾80aa) threshold was 1188 in maize, 1287 in soybean, 830 in *O.s.* ssp. *indica* and 1097 in *O.s.* ssp. *japonica* (Table 1).

The genomes of *G. max*, *Z. mays*, and *O. sativa* all contain naturally occurring allergens listed in the FARRP 11 database. *G. max* had the highest number of allergens with 36, followed by *Z. mays* at 24, and *O. sativa* with 20 confirmed FARRP allergens. To demonstrate that these endogenous allergens do not account for a substantial fraction of the novel polypeptides alignments exceeding the Codex threshold, each species-specific allergen was screened for in the top scoring alignments from each respective genome dataset. Only a very small fraction of the top alignments were to an endogenous allergen (Table 2). The largest number occurred in *G. max*, where 171 (∼5%) of the top alignments were to known endogenous soybean allergens listed in the FARRP11 database, whereas in maize less than 1% of the novel polypeptides had an endogenous maize allergen as its top-scoring alignment. Therefore, the FARRP 11 allergens endogenous to each of the genomes had little effect (less than 5%) on the total number of novel polypeptides exceeding the Codex's 35%/⩾80aa threshold.

3.2. E-value and distribution of the matches to the FARRP 11 database

The expected, or *E*-value, is commonly used in BLAST or FASTA as statistical measurement of the likelihood that an alignment is due to chance in the database searched. A low *E*-value signifies a lower likelihood that an alignment is due to chance and a higher likelihood of true sequence homology. The Codex, however, does not provide any guidance on the *E*-value of an alignment as part of the bioinformatics assessment. As such, alignments that exceed the 35%/⩾80aa threshold, but have very poor statistical significance, are still considered a potential cross-reactivity risk under the guidelines. Such an approach is known to lead to exceedingly high false positive rates for 35%/⩾80aa matches (Silvanovich et al., 2009). To characterize the *E*-value distribution of the novel polypeptides matches to the FARRP 11 database, novel polypeptides were grouped by the *E*-value of their top scoring alignment that exceeded the Codex 35%/⩾80aa threshold in the FASTA analysis (Fig. 4). The results show that approximately 47% (37,439 novel polypeptides) of the top scoring alignments had an *E*-value above 0.01. In other words, the alignment could be expected to occur by chance in at least one of every one hundred searches or less (Pearson, 2000). Moreover, maize contained 1927 novel polypeptides with an *E*-value above one which indicated that the alignment was expected to occur by chance alone and that no true sequence homology existed. Nonetheless, the Codex guidelines state that such instances may require additional testing because the *E*-value is not part of the sequence identity assessment. For the other 42,910 novel polypeptides for which the top-scoring alignment had an *E*-value below 0.01, a substantial fraction (60%) show some moderate level of significance, given the small size of the FARRP 11 database (1491 proteins), with scores between 0.0001 and 0.01. A total of 5699 of the novel polypeptides from all five genomes had an *E*-value below 1×10^{-10}, suggesting that at least some sequence homology exists with the allergens in the database (Table S2). A set of highly significant matches (*E*-value less than 1×10^{-20}) was cross-referenced to the set of novel polypeptides with matches to protein-coding sequence. We observed that the majority of those alignments with highly significant *E*-values were from novel polypeptides that matched protein-coding sequence (76% for *G. max*, Z. *mays*, *japonica*, and 63% for *indica*).

The frequency of each allergen across the top scoring alignment for each novel polypeptide that exceeded the 35%/⩾80aa threshold was counted in order to evaluate whether there were certain allergens responsible for a disproportionate number of matches. The top three allergen matches are shown in Table 3. The same three allergens, Bos collagen alpha2, Lol p 5, and Amb 4 defensin-like protein were observed for the maize, rice, and human datasets. The Bos collagen alpha2 protein from cow was the most frequent match in four of the five datasets. Maize alone had 13,532 matches to this one allergen, comprising 40% of the total number of matches for maize, while a total of 11,737 matches were observed in the human dataset. *G. max* was the only exception where this pattern was not observed. The Bos collagen alpha2 protein is the alpha2

chain of the bovine type 1 collagen protein. It has been found as the dominant IgE binding protein in patients with gelatin allergies (Sakaguchi et al., 1999). The majority of matches to this protein have high *E*-values (>0.01) suggesting they are not homology based even though they exceed the 35%/⩾80aa threshold (Table S3). Moreover, the Bos collagen alpha 2 protein is 1364aa long and contains large numbers of glycine and proline residues, 27.8% and 17.3% of the total aa content respectively, suggesting that the above threshold alignment is likely the result of locally biased amino acid composition (GenBank ID: 27806257). Similarly, heavily biased composition is also observed for Lol p 5 (ryegrass) and the Amb 4 defensin-like protein (ragweed).

Despite the prevalence of high frequency alignments (e.g. Bos collagen alpha2, Lol p 5, and Amb 4 defensin-like protein), the set of novel polypeptides match a large number of different allergens from the FARRP 11 database. For example, the set of novel polypeptides from maize returned matches to 552 different allergens in the database. Soybean, rice *japonica* and *indica* varieties, and human derived novel polypeptides matched 361, 382, 418, and 465 different allergens, respectively. The fraction of matches to a particular allergen group and the fraction each group represents in the FARRP 11 database are shown in Fig. 5 and Table S5. The plant allergen group is by far the largest group of allergens in the FARRP 11 database; this group also comprises the largest fraction of matches to the novel polypeptide sets from the plant genomes. There were fewer matches to the animal and insect groups, with most of the matches observed in the maize dataset (172 animal and 174 insect allergen matches), while very few were observed in the other three plant datasets. The "Other" group consists of allergens derived from bacteria, protozoa, and nematodes, as well as contact allergens, celiac proteins (gliadins/glutelins), vaccine-based allergens, and unassigned sequences. The fraction of matches to the 'Other' group is over-represented by the matches to the Bos collagen alpha2 protein, which is considered a vaccine related allergen, with the exception of the soybean dataset which has only 80 matches to Bos collagen alpha2 (Tables S3 and S4).

3.3. Eight residue identical matches to the FARRP 11 database

In addition to assessing the number of stop-to-stop frames exceeding the 35%/⩾80aa threshold, the number of eight residue identical matches (hereafter referred to as 8-mers) to the FARRP 11 database was also assessed. The allergen database contained hundreds of thousands of matches, ranging from 71,332 in *O.s.* ssp. *japonica* to 217,874 in *Z. mays*. A total of 234,419 matches were observed within the human dataset. The most common 8-mer matches consisted of low complexity sequence. For example, the 'SSSSSSSS' 8-mer (8 serine residues) occurs 46,403, 9255, 9698, 9408, and 38,540 times in *Z. mays*, *G. max*, *O.s. japonica* and *indica*, and the human datasets, respectively (Table 5). The 'SSSSSSSS' 8-mer appears in only three allergens from the FARRP 11 database, each of which is from wheat. Other common 8-mer matches are summarized in Table 5. The number of unique 8-mers with one or more identical matches to the database ranged from approximately 12,000 in *japonica* to nearly 32,000 in humans. Similarly, we observed thousands of unique 8-mers from the FARRP11 database with one or more identical matches to the set of protein-coding genes from each genome (Table 4).

3.4. Intraspecies comparison reveals diversity of novel polypeptides

We estimated the number of novel polypeptides conserved between rice *O.s.* ssp. *japonica* and *indica* varieties using BLASTp. Each set of novel polypeptides was used as a query to the other set and all alignments above 95% identity across a length of at least 80% of the query sequence were scored as being conserved (shared) between the two rice varieties and reported in Table 6. We observed more than six thousand novel polypeptides in *japonica* and more than five thousand in *indica*, more than half of the total number of novel polypeptides from each genome that exceeded the 35%/⩾80aa threshold, that were unique. Dropping the percent query length down to 50% only marginally increases the number of conserved novel polypeptides. For example, in the *japonica* versus *indica* comparison the number of unique novel polypeptides at the 50% query length threshold increases to 5308, a difference of only 410 sequences. In the reciprocal comparison (*indica* versus *japonica*), the number increases to 4953, a difference of only 381 sequences. Some of the differences between *indica* and *japonica* may be attributable to the quality and completeness of the genome datasets (Ouyang et al., 2007; Yu et al., 2002). Nonetheless, we estimate that approximately half of the novel polypeptides found translated from the *japonica* and *indica* genomes are unique to one variety.

To further illustrate natural variation, we analyzed four well-characterized loci from the maize inbred lines B73 and Mo17 in a similar manner as was performed for the genome analysis. The following four loci were selected: 9002, 9008, 9009, and *bz1*. Each locus was characterized in Brunner et al., 2005 and was shown to contain high levels of variation, primarily due to the proliferation of retrotransposons. A single BAC sequence covered the entire

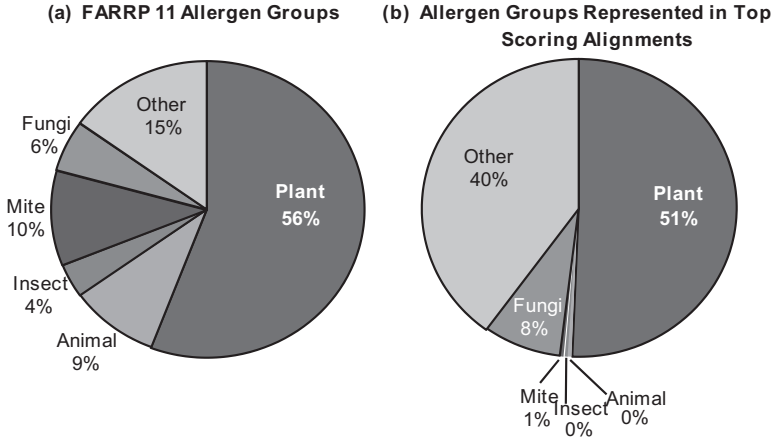

Fig. 5. (a) Percentage each allergen group represents in the FARRP 11 database. (b) Total percentage each allergen group represents in all top scoring alignments from all four plant genomes analyzed combined.

Table 5
The ten most frequent 8-mer matches to the FARRP 11 database per genome dataset.

G. max		*Indica*[a]		*Japonica*[b]		*Z. mays*		*H. sapiens*	
SSSSSSSS	9255	SSSSSSSS	9698	SSSSSSSS	9408	SSSSSSSS	46403	SSSSSSSS	38540
EEEEEEEE	4351	EEEEEEEE	3255	EEEEEEEE	3361	QQQQQQQQ	7613	EEEEEEEE	25198
QQQQQQQQ	2825	SSSSSSSL	1353	TPPPPPPP	1218	SSSSSSSL	3032	QQQQQQQQ	11095
LSFILTYF	2214	QQQQQQQQ	1145	QQQQQQQQ	1016	SSSSSGVS	2755	KKKKKKTK	5374
ERERERES	905	TPPPPPPP	1121	ERERERES	903	LVPSGLIL	2726	KERERERE	5331
KERERERE	864	ERERERES	917	VAAAAAAA	897	SEESEDSE	2247	KQQQQQQQ	3778
TPPPPPPP	812	VAAAAAAA	897	GGGGGGEG	831	SGGAGGAS	2225	ERERERES	2893
TSSSSSSS	801	LNRNNSFK	716	SSSSSSSL	812	TSSSSSSS	2076	QQQQQQQK	2482
SSSSSSSL	796	GGGGGGEG	712	RSSSSSSS	685	SLEGELKG	1995	LPLLLLLL	1958
LSYLKKGG	691	TSSSSSSS	650	TSSSSSSS	643	EEEEEEEE	1939	ERERQRER	1786

[a] *O.s.* ssp. *indica*.
[b] *O.s.* ssp. *japonica*.

Table 6
Blastp comparison of the novel polypeptides exceeding the Codex 35%/⩾80aa from *O.s.* ssp. *japonica* and *indica*.

Query vs. database[a]	Shared [b]	Non-shared [c]
Japonica vs. *indica*	4898	6581
Indica vs. *japonica*	4572	5462

[a] The *japonica* set was used as a query against the database of *indica* novel polypeptides and vice versa.
[b] The cutoff for shared was 95% identity across an alignment of at least 80% of the length of the query sequence.
[c] The number of novel polypeptides that fall below the 95% identity across 80% of the query length.

region analyzed for each locus. Across the approximately 2 MB of sequence that was analyzed, we discovered a total of 83 novel polypeptides with alignments to an allergen above the Codex threshold that were unique to either B73 or Mo17, 73 of which were clearly derived within a retrotransposon sequence (Table 7). Almost all of the 38 novel polypeptides that were conserved between B73 and Mo17, and exceeded the 35%/⩾80aa threshold, spanned a coding sequence. Moreover, to investigate what fraction of the total number of novel polypeptide from the maize genome dataset were derived from transposable elements and repetitive sequence, we performed a tBLASTn analysis to the maize repeat database (maize.jcvi.org.repeat_db.shtml, version 4.0) using the subset of novel polypeptides that had an alignment above the 35%/⩾80aa threshold, but did not have a match to protein-coding sequence, as the query sequence. A total of 5558 novel polypeptides had matches to one or more entries in the repeat database above a score of 90% identity across an alignment that was at least 80% of the length of the query/novel polypeptide sequence. This set of 5558 novel polypeptides match 181 different entries in the FARRP 11 database, which shows that they were not specific to a handful or fewer allergens. These results also suggest that a substantial fraction (~16%) of the novel polypeptides exceeding the Codex threshold from the maize genome are derived from transposable and repetitive sequence, which are known to be highly variable across different lines of maize (Brunner et al., 2005; Wang and Dooner, 2006; Morgante et al., 2005).

4. Discussion

4.1. Assessment of allergenicity of common food crop genomes using the Codex 35%/⩾80aa guideline

In this study, the same bioinformatics safety assessment required for transgenic insertions (see Fig. 1) was applied across the reference genomes of common, non-transgenic crops such as maize, soybean, and rice. Our results identified thousands of stop-to-stop reading frames (novel polypeptides) within the

Table 7
Comparison of the number of shared and non-shared novel polypeptides between maize inbred lines B73 and Mo17 across four well-characterized loci.

			Novel polypeptides				
Locus	GB accession	Length (bp)	Total ⩾80aa[a]	35%/⩾80aa[b]	Shared[c]	Non-shared[d]	Retrotransposons[e]
9002							
B73	AY664413	317,137	1122	19 (1.7%)	4	15	11
Mo17	AY664417	366,120	1337	25 (1.9%)	4	21	14
9008							
B73	AY664414	339,089	1146	14 (1.2%)	3	11	10
Mo17	AY664418	282,600	987	11 (1.1%)	5	6	9
9009							
B73	AY664415	323,584	1101	12 (1.1%)	10	2	4
Mo17	AY664419	405,672	1339	15 (1.1%)	9	6	7
bz1							
B73	AF448416	106,186	343	20 (5.8%)	1	19	17
Mo17	AY664416	203,581[f]	726	15 (2.1%)	2	3	1
Total		2,343,969	8101	131(1.6%)	38	83	73

[a] The number of novel polypeptides ⩾80aa in length.
[b] The number of novel polypeptides that are ⩾80aa in length and have at least one match to the FARRP 11 database that exceeds the Codex 35%/⩾80aa threshold.
[c] The number of shared novel polypeptides between B73 and Mo17 with a 35%/⩾80aa match to the FARRP 11 database.
[d] The number of non-shared novel polypeptides between B73 and Mo17 with a 35%/⩾80aa match to the FARRP 11 database.
[e] The number of 35%/⩾80aa novel polypeptides derived from retrotransposon sequence.
[f] The range of comparison at the *bz1* locus between B73 and Mo17 was limited to the region for which the two loci had homologous sequence; only those novel polypeptides between nucleotide position 1 and 94,236 in AY664416 were compared against B73 (Brunner et al., 2005).

genomic sequence that had one or more alignments which exceeded the 35%/⩾80aa threshold set by the Codex (2009) guidelines for presenting a potential cross-reactivity risk (Table 1). Such a large number of frames exceeding the 35%/⩾80aa threshold suggest that each of these crops would fail the bioinformatic safety assessment as it is applied to GM crops. If such novel polypeptides (translated stop-to-stop frames) represented an unacceptable safety risk, then incidences of cross-reactivity from these conventional non-transgenic crops could be expected. For example, we identified a total of 210 novel polypeptides from soybean for which the top scoring alignment exceeding 35%/⩾80aa was to Cuc m 1, the major food allergen of muskmelon, yet there are no documented cases of cross-reactivity between soybean and the Cuc m 1 allergen (Table 3). Nor are there any documented cases of cross-reactivity between maize and peanut-allergic patients despite over 400 maize-derived novel polypeptides for which the top scoring alignment exceeding 35%/⩾80aa was to one or more of the forms of the major peanut allergen Ara h 1 (data not shown). Other similar examples, which are shown in Supplementary Table S5, would include 84 and 574 top-scoring alignments from soybean and maize respectively to Para rubber tree (natural latex) allergens, and 47 and 156 top-scoring alignments to the mosquito allergen Aed a 3 from soybean and rice, respectively. These examples, in addition to the thousands of other novel polypeptides with alignments that exceeded the Codex threshold, suggest that sequences derived from stop-to-stop frames are unlikely to represent an unacceptable allergen risk because such a picture is not consistent with the safe history and what is generally known about the protein families that contain allergens from these food crops (Breiteneder and Radauer, 2004; Radauer et al., 2008). Rather, the risk presented by such sequences is largely, if not all, theoretical, especially considering that hypothetical ORFs, such as the stop-to-stop frames analyzed here, do not contain the necessary cis sequence elements typically associated with stable gene expression (start codon, promoter sequence, terminator sequence, intron splice sites, ribosomal binding site, etc.). Moreover, many of the ORFs represent alternative frames of existing protein-coding genes. Nuclear encoded protein-coding sequences in higher eukaryotes, such as plants, are not known to overlap in such a way.

The results from this study also further highlight the large number of seemingly false positive matches produced by the conservative 35%/⩾80aa criteria (Goodman, 2008; Goodman et al., 2008; Cressman and Ladics, 2009; Silvanovich et al., 2009; Ladics et al., 2011; Harper et al., 2012). Only *G. max* (soybean) is considered a major source for food allergies, whereas maize and rice are only minor sources (Hefle et al., 1996; Moneret-Vautrin et al., 1998), yet soybean contained the fewest number of novel polypeptides exceeding the Codex threshold of 35%/⩾80aa despite containing a similar overall number of novel polypeptides as either of the two rice varieties. Moreover, the human genome, which we included to serve as a large, non-plant, non-allergen source control, contained well over 20,000 unique novel polypeptides that had one or more alignments that exceeded the Codex threshold. When looking at protein-coding genes, this analysis likely missed many proteins that may have exceeded the Codex criteria, but did not have an alignment greater than 80% of the length of a novel polypeptide sequence. For example, the Harper et al. (2012) study, which compared translated stop-to-stop frames from annotated genes from a variety of food plants to the FARRP11 database using FASTA, identified more than 7000 and 10,000 unique genes with alignments that exceeded the Codex threshold from maize and rice respectively, whereas our analysis only identified approximately a thousand from maize and rice. Harper et al. also identified thousands of alignments that exceeded the Codex threshold from other common foods such as cucumber, melon, watermelon, and tomato that exceeded the threshold. Such results from the Harper et al. (2012) study and this study, suggest that the 35%/⩾80aa criteria needs to be improved if it is to be considered a reliable predictor of cross-reactivity. Allowing additional criteria, such as an *E*-value threshold, may be a simple way to add value by improving the specificity (i.e. lowering the number false positive alignments) with little to no risk of missing true allergens (i.e. false negatives) (Cressman and Ladics, 2009; Silvanovich et al., 2009). For example, close to half of the top scoring alignments from all 80,349 novel polypeptides that exceeded the 35%/⩾80aa threshold had an *E*-value above 0.01, which is widely considered a non-significant score. When using what may be considered a conservative *E*-value threshold of 3.9×10^{-7} as recommended by Silvanovich et al. (2009), the number of unique novel polypeptides for which the top scoring alignment exceeded the 35%/⩾80aa threshold reported in Table 1 would drop to 2197, 2036, 1326, 1365, and 1769 for maize, soybean, rice varieties *japonica* and *indica*, and human, respectively. This would represent more than a 10-fold reduction for the number of novel polypeptides in the maize and human data sets, an approximate 10-fold reduction for the two rice varieties (*japonica* and *indica*), while only a reduction of approximately 1300 novel polypeptides in soybean.

Another issue with current bioinformatics guidelines is the lack of any guidance on how to contend with low-complexity or compositionally biased sequences. Low complexity sequence violates the underlying assumptions in the FASTA algorithm and will artificially inflate the number of matches with significant *E*-values or local sequence identity such as the 35% identity over 80 or more amino acids. The assessment of sequence homology using the FASTA default settings does not allow for the masking of low-complexity sequence in either the query or database. The large number of matches to the glycine/proline rich Bos collagen alpha2 protein is a good example of how biased amino acid composition can inflate the number of alignments. Out of the 13,532 novel polypeptides found in maize for which the top scoring alignment was to the Bos collagen alpha2 protein, only one alignment had an *E*-value below 1×10^{-20} (Table S3). The query sequence producing this alignment is 464aa in length, of which 69% (319aa) are glycine residues. As a result, the alignment produced by FASTA for this 464aa novel polypeptide and the Bos collagen alpha2 protein was significant because it stretched across 444aa, but is not biologically significant in terms of establishing homology. This underscores the need for additional guidance from regulatory bodies on how to analyze compositional biased or repetitive sequences, which may be common within the flanking sequence and/or the insert/genomic junctions. Considering the *E*-value of an alignment would help eliminate some compositional biased sequences that exceed the local similarity threshold of 35%/⩾80aa. Though, it is likely that additional measures such as using low-complexity filtering algorithms would be required to help distinguish between homology based alignments and those due to biased or low complexity sequence.

4.2. Short contiguous identical matches do not provide additional value to the allergenicity assessment

Short contiguous identical segments of eight or more amino acids could represent theoretical IgE binding epitopes and therefore are currently evaluated in the allergenicity assessment independent of the 35%/⩾80aa analysis. Performing this type of search on all the translated stop-to-stop frames eight or more amino acids in length using the minimum length generated hundreds of thousands of matches across each of the five genome datasets. In addition, a separate search using the annotated protein-coding sequence showed that thousands of matches to the allergen database are present within these expressed sequences (Table 4). Since the search for short continuous identical matches was limited only to

eight residues in length, multiple 8-mer matches may have been derived from a single novel polypeptide or peptide sequence containing a long high-identity region or multiple high-identity regions. However, the Harper et al. (2012) reported thousands of unique genes from common food crops with 8-mer matches to the FARRP11 allergen database. Such data suggests that the 8-mer matches are distributed widely across the genome despite the safe history and widespread use of these common food crops. As a result the 8-mer matches are likely false-positives (with the exception of those from endogenous allergens) and therefore are not reliable for predicting the cross-reactivity risk of a particular hypothetical ORF or newly expressed protein. Experimental evidence would suggest that for an 8-mer to result in serum IgE cross-reactivity it would have to be in the same conformational context as in the allergen (i.e. there would need to be significant overall sequence homology between the newly expressed protein and the allergen) (Klinglmayr et al., 2009). For example, in the Klinglmayr et al., (2009) study, which identified the possible epitopes involved in the cross-reactivity of Bet v 1 IgE antibodies with the major allergen from apple (Mal d 1), the two proteins (Bet v 1 and Mal d 1) share more than 64% amino acid identity and are conserved in structure. Therefore, it is unclear whether any additional value is gained by analyzing short contiguous identical segments of 8 residues in length—which without empirical data represent only theoretical epitopes—independent of an overall sequence similarity search, such as the 35%/⩾80aa criteria.

4.3. Intraspecies variation in novel polypeptide content demonstrates that stop-to-stop frames are unlikely to pose an unacceptable safety risk

The genomes of the common food crops are not static, but instead are plastic and contain significant amounts of variation. More than half of the identified novel polypeptides that exceeded the Codex 35%/⩾80aa guideline threshold for rice were unique to either the *japonica* or *indica* rice variety (Table 6). Despite this variability, it is commonly assumed there is no novel allergenicity risk associated with either variety, which is supported by their history of safe use. If hypothetical ORFs (stop-to-stop frames) represented an unacceptable cross-reactivity or allergen risk, then variability in sequence content would be expected to cause differences in allergenicity independent of variation in endogenous allergen expression levels. Since this is generally considered not to be the case, the analysis of hypothetical ORFs remains a theoretical exercise and thus its utility in evaluating the safety of a transgene insertion is questionable. Furthermore, the comparison across the four loci from maize lines B73 and Mo17 showed that retrotransposons contained stop-to-stop frames that exceeded the 35%/⩾80aa threshold. This is especially significant when considering that approximately 84% of the maize genome is estimated to be transposable elements (Schnable et al., 2009), which may help explain why maize contained the highest number of novel polypeptides that exceeded the guidelines. For example, greater than 5000 of the 35%/⩾80aa novel polypeptide set from maize had a match to the maize repeat database. Such repetitive content is known to be highly variable within maize (Brunner et al., 2005; Wang and Dooner, 2006; Morgante et al., 2005) and rice (Ammiraju et al., 2008). There are also documented cases of differential accumulation of retroelements within soybean (Innes et al., 2008). In addition, the *Helitron* transposons in maize are known to frequently contain fragments of protein-coding genes and contribute to the genome structure heterogeneity observed across different maize lines (Lai et al., 2005; Du et al., 2009; Wang and Dooner, 2006). Such instances, especially the movement of *Helitrons* containing gene fragments, can lead to chimeric frames analogous to the insert/genomic junction at the insertion site of a transgene. If there is an acceptable risk generally associated with naturally occurring changes from the movement of transposable elements or gene fragments in conventional, non-transgenic crops, then it is unclear what criteria distinguishes the concerns surrounding the changes created by the insertion of a single transgene in a GM crop?

5. Conclusion

Bioinformatic analysis can be extremely helpful in identifying similarities between sequences and is a good first step for assessing potential allergenicity of a newly expressed protein. However, it alone cannot reliably predict whether or not a protein will act as an allergen or cross-reactive with one. Ultimately, empirical data may be needed to rule out potential for cross-reactivity. Bioinformatics is used as only a single step in weight of evidence approach outlined in the Codex guidelines and can be valuable in determining whether certain proteins warrant additional tests. Some regulatory authorities have broadened the scope of the Codex (2009) bioinformatic guidelines on newly expressed proteins to include all sequences between stop codons from all six DNA reading frames present within the insert and spanning the insert/genomic DNA junction sites with no minimum length. Applying this requirement for GM crops to the reference genomes of maize, soybean, and two rice varieties showed that these non-transgenic common food crops would fail this bioinformatic safety assessment despite their safe history and widespread consumption. As a result, the requirement to evaluate all stop-to-stop codons without any consideration for potential gene expression has questionable value in terms of making decisions on the safety and testing of GM crops. This is especially true, when considering that thousands of stop-to-stop frames may be created/introduced through unregulated conventional breeding practices or the movement of transposable elements. To help minimize the testing of hypothetical ORFs the analysis of stop-to-stop frames could be restricted to just those frames spanning the insert/genomic junction. If there is evidence that suggests a novel fusion protein was created by the insertion event, either within the transgene or spanning an insert/genomic junction, then such ORFs should be evaluated according to Codex guidelines. In addition, considering criteria such as *E*-value would be a simple and useful way to increase the specificity of the bioinformatic assessment for newly expressed proteins or hypothetical ORFs (Cressman and Ladics, 2009; Silvanovich et al., 2009). However, including such requirements as searching for short continuous identical matches as small as eight residues in length does not appear to add any additional value to the safety assessment given that so many matches occur naturally and are unlikely to present a cross-reactivity risk.

Conflict of Interest

All authors are employed by Pioneer Hi-Bred International, Inc., a DuPont Company.

Acknowledgements

The authors would like to thank Natalie Weber, Annie Gutsche, Wim Broothaerts, Ray Layton, Antoni Rafalski, and Mary Locke for their critical reading and comments on the manuscript.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.fct.2012.07.044.

References

Ammiraju, J.S., Lu, F., Sanyal, A., Yu, Y., Song, X., Jiang, N., Pontaroli, A.C., Rambo, T., Currie, J., Collura, K., Talag, J., Fan, C., Goicoechea, J.L., Zuccolo, A., Chen, J., Bennetzen, J.L., Chen, M., Jackson, S., Wing, R.A., 2008. Dynamic evolution of oryza genomes is revealed by comparative genomic analysis of a genus-wide vertical data set. Plant Cell 20, 3191–3209.

Breiteneder, H., Radauer, C., 2004. A classification of plant food allergens. J. Allergy Clin. Immunol. 113, 821–830.

Brunner, S., Fengler, K., Morgante, M., Tingey, S., Rafalski, A., 2005. Evolution of DNA Sequence Nonhomologies among Maize Inbreds. Plant Cell 17, 343–360.

Codex Alimentarius Commission, 2009. Foods Derived From Modern Biotechnology. FAO/WHO, Rome, pp. 1–85.

Cressman, R.F., Ladics, G.S., 2009. Further evaluation of the utility of "Sliding Window" FASTA in predicting cross-reactivity with allergenic proteins. Regul. Toxicol. Pharmcol. 54, S20–S25.

Du, C., Fefelova, N., Caronna, J., He, L., Dooner, H.K., 2009. The polychromatic *Helitron* landscape of the maize genome. Proc. Natl. Acad. Sci. USA 106, 19916–19921.

EFSA, 2011. EFSA panel on genetically modified organisms (GMO). Scientific opinion on guidance for risk assessment of food and feed from genetically modified plants. EFSA J. 9(5), 2150 [37p].

Goodman, R.E., 2008. Performing IgE serum testing due to bioinformatics matches in the allergenicity assessment of GM crops. Food Chem. Toxicol. 46, S24–S34.

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. Biotech. 26, 73–81.

Harper, D., McClain, S., Ganko, E.W., 2012. Interpreting the biological relevance of the bioinformatic analysis with T-DNA sequence for protein allergenicity. Regul. Toxicol. Pharmcol. 63, 426–432.

Hefle, S.L., Nordlee, J.A., Taylor, S.L., 1996. Allergenic foods. Crit. Rev. Food Sci. Nutr. 36, 69–89.

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.

Innes, R.W., Ameline-Torregrosa, C., Ashfield, T., Cannon, E., Cannon, S.B., Chacko, B., Chen, N.W., Couloux, A., Dalwani, A., Denny, R., Deshpande, S., Egan, A.N., Glover, N., Hans, C.S., Howell, S., Ilut, D., Jackson, S., Lai, H., Mammadov, J., Del Campo, S.M., Metcalf, M., Nguyen, A., O'Bleness, M., Pfeil, B.E., Podicheti, R., Ratnaparkhe, M.B., Samain, S., Sanders, I., Ségurens, B., Sévignac, M., Sherman-Broyles, S., Thareau, V., Tucker, D.M., Walling, J., Wawrzynski, A., Yi, J., Doyle, J.J., Geffroy, V., Roe, B.A., Maroof, M.A., Young, N.D., 2008. Differential accumulation of retroelements and diversification of NB-LRR disease resistance genes in duplicated regions following polyploidy in the ancestor of soybean. Plant Physiol. 148, 1740–1759.

International Rice Genome Sequencing Project, 2005. The map-based sequence of the rice genome. Nature 436, 793–800.

James, C., 2011. Global Status of Commercialized Biotech/GM Crops, ISAAA Brief No. 43.

Klinglmayr, E., Hauser, M., Zimmermann, F., Dissertori, O., Lackner, P., Wopfner, N., Ferreira, F., Wallner, M., 2009. Identification of B-cell epitopes of Bet v 1 involved in cross-reactivity with food allergens. Allergy 64, 647–651.

Ladics, G.S., Bardina, L., Cressman, R.F., Mattsson, J.L., Sampson, H.A., 2006. Lack of cross-reactivity between the *Bacillus thuringiensis* derived protein Cry1F in maize grain and dust mite Der p7 protein with human sera positive for Der p7-IgE. Regul. Toxicol. Pharmacol. 44, 136–143.

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.

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.

Lai, J., Li, Y., Messing, J., Dooner, H.K., 2005. Gene movement by Helitron transposons contributes to the haplotype variability of maize. Proc. Natl. Acad. Sci. USA 102, 9068–9073.

Moneret-Vautrin, D.A., Kanny, G., Beaudouin, E., 1998. L'allergie alimentaire au mais existe-t-elle? (Does corn allergy exists?). Allergy Immunol. 30, 230.

Morgante, M., Brunner, S., Pea, G., Fengler, K., Zuccolo, A., Rafalski, A., 2005. Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. Nat. Genet. 37, 997–1002.

Ouyang, S., Zhu, W., Hamilton, J., Lin, H., Campbell, M., Childs, K., Thibaud-Nissen, F., Malek, R.L., Lee, Y., Zheng, L., Orvis, J., Haas, B., Wortman, J., Buell, C.R., 2007. The TIGR Rice Genome Annotation Resource. improvements and new features. Nucleic Acids Res. 35, D883–D887.

Pearson, W.R., Lipman, D.J., 1988. Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. USA 85, 2444–2448.

Pearson, W.R., 2000. Flexible sequence similarity searching with the FASTA3 program package. Methods Mol. Biol. 132, 185–219.

Radauer, C., Bublin, M., Wagner, S., Mari, A., Breiteneder, H., 2008. Allergens are distributed into few protein families and possess a restricted number of biochemical functions. J. Allergy Clin. Immunol. 121, 847–852.

Rice, P., Longden, I., Bleasby, A., 2000. EMBOSS: The European Molecular Biology Open Software Suite. Trends Genet. 16, 276–277.

Sakaguchi, M., Hori, H., Hattori, S., Irie, S., Imai, A., Yanagida, M., Miyazawa, H., Toda, M., Inouye, S., 1999. IgE reactivity to alpha1 and alpha2 chains of bovine type 1 collagen in children with bovine gelatin allergy. J. Allergy Clin. Immunol. 104, 695–699.

Stephen, F.A., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402.

Schmutz, J., Cannon, S.B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., Hyten, D.L., Song, Q., Thelen, J.J., Cheng, J., Xu, D., Hellsten, U., May, G.D., Yu, Y., Sakurai, T., Umezawa, T., Bhattacharyya, M.K., Sandhu, D., Valliyodan, B., Lindquist, E., Peto, M., Grant, D., Shu, S., Goodstein, D., Barry, K., Futrell-Griggs, M., Abernathy, B., Du, J., Tian, Z., Zhu, L., Gill, N., Joshi, T., Libault, M., Sethuraman, A., Zhang, X.-C., Shinozaki, K., Nguyen, H.T., Wing, R.A., Cregan, P., Specht, J., Grimwood, J., Rokhsar, D., Stacey, G., Shoemaker, R.C., Jackson, S.A., 2010. Genome sequence of the palaeopolyploid soybean. Nature 463, 178–183.

Schnable, P.S., Ware, D., Fulton, R.S., Stein, J.C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T.A., Minx, P., Reily, A.D., Courtney, L., Kruchowski, S.S., Tomlinson, C., Strong, C., Delehaunty, K., Fronick, C., Courtney, B., Rock, S.M., Belter, E., Du, F., Kim, K., Abbott, R.M., Cotton, M., Levy, A., Marchetto, P., Ochoa, K., Jackson, S.M., Gillam, B., Chen, W., Yan, L., Higginbotham, J., Cardenas, M., Waligorski, J., Applebaum, E., Phelps, L., Falcone, J., Kanchi, K., Thane, T., Scimone, A., Thane, N., Henke, J., Wang, T., Ruppert, J., Shah, N., Rotter, K., Hodges, J., Ingenthron, E., Cordes, M., Kohlberg, S., Sgro, J., Delgado, B., Mead, K., Chinwalla, A., Leonard, S., Crouse, K., Collura, K., Kudrna, D., Currie, J., He, R., Angelova, A., Rajasekar, S., Mueller, T., Lomeli, R., Scara, G., Ko, A., Delaney, K., Wissotski, M., Lopez, G., Campos, D., Braidotti, M., Ashley, E., Golser, W., Kim, H., Lee, S., Lin, J., Dujmic, Z., Kim, W., Talag, J., Zuccolo, A., Fan, C., Sebastian, A., Kramer, M., Spiegel, L., Nascimento, L., Zutavern, T., Miller, B., Ambroise, C., Muller, S., Spooner, W., Narechania, A., Ren, L., Wei, S., Kumari, S., Faga, B., Levy, M.J., McMahan, L., Van Buren, P., Vaughn, M.W., Ying, K., Yeh, C.-T., Emrich, S.J., Jia, Y., Kalyanaraman, A., Hsia, A.-P., Barbazuk, W.B., Baucom, R.S., Brutnell, T.P., Carpita, N.C., Chaparro, C., Chia, J.-M., Deragon, J.-M., Estill, J.C., Fu, Y., Jeddeloh, J.A., Han, Y., Lee, H., Li, P., Lisch, D.R., Liu, S., Liu, Z., Nagel, D.H., McCann, M.C., Sanmiguel, P., Myers, A.M., Nettleton, D., Nguyen, J., Penning, B.W., Ponnala, L., Schneider, K.L., Schwartz, D.C., Sharma, A., Soderlund, C., Springer, N.M., Sun, Q., Wang, H., Waterman, M., Westerman, R., Wolfgruber, T.K., Yang, L., Yu, Y., Zhang, L., Zhou, S., Zhu, Q., Bennetzen, J.L., Dawe, R.K., Jiang, J., Jiang, N., Presting, G.G., Wessler, S.R., Aluru, S., Martienssen, R.A., Clifton, S.W., McCombie, W.R., Wing, R.A., Wilson, R.K., 2009. The B73 maize genome: complexity, diversity, and dynamics. Science 326, 1112–1115.

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.

Silvanovich, A., Bannon, G., McClain, S., 2009. The use of E-scores to determine the quality of protein alignments. Regul. Toxicol. Pharmacol. 54, S26–S31.

Stadler, M.B., Stadler, B.M., 2003. Allergenicity prediction by protein sequence. FASEB J. 17, 1141–1143.

Wang, Q., Dooner, H.K., 2006. Remarkable variation in maize genome structure inferred from haplotype diversity at the bz locus. Proc. Natl. Acad. Sci. USA 103, 17644–17649.

Yu, J., Hu, S., Wang, J., Wong, G.K.-S., Li, S., Liu, B., Deng, Y., Dai, L., Zhou, Y., Zhang, X., Cao, M., Liu, J., Sun, J., Tang, J., Chen, Y., Huang, X., Lin, W., Ye, C., Tong, W., Cong, L., Geng, J., Han, Y., Li, L., Li, W., Hu, G., Li, J., Liu, Z., Qi, Q., Li, T., Wang, X., Lu, H., Wu, T., Zhu, M., Ni, P., Han, H., Dong, W., Ren, X., Feng, X., Cui, P., Li, X., Wang, H., Xu, X., Zhai, W., Xu, Z., Zhang, J., He, S., Xu, J., Zhang, K., Zheng, X., Dong, J., Zeng, W., Tao, L., Ye, J., Tan, J., Chen, X., He, J., Liu, D., Tian, W., Tian, C., Xia, H., Bao, Q., Li, G., Gao, H., Cao, T., Zhao, W., Li, P., Chen, W., Zhang, Y., Hu, J., Liu, S., Yang, J., Zhang, G., Xiong, Y., Li, Z., Mao, L., Zhou, C., Zhu, Z., Chen, R., Hao, B., Zheng, W., Chen, S., Guo, W., Tao, M., Zhu, L., Yuan, L., Yang, H., 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). Science 296, 79–92.